Introduction to Applied Statistics.
A Modelling Approach
Instructor's Notes

J.K. Lindsey
Department of Social Sciences,
Université de Liège
jlindsey@luc.ac.be

February 16, 2020

Preface

I offer these notes as a rather personal description of my experiences with giving this course over the past 28 years, introducing statistics to non-mathematics/statistics majors.

I would like to thank Muriel Comblain, Marie-Hélène Delsemme, and Philippe Lambert for their contributions to the elaboration of these notes, as well as the feedback and enthusiasm of all of my students over the years.

Diepenbeek and Liège J.K.L.
January, 2004

Contents

Chapter 1

Basic concepts

1.1 Variables

1.1.1 Definition

Emphasise that the first lecture, on variable construction (Section 1.1), may be the most important of the whole course. Without valid construction of variables, nothing else will be possible as far as statistical methods are concerned. Point out that this is more generally true in any scientific endeavour and that the techniques for variable construction that they are learning will be more widely applicable than just to statistics. Non-statistical approaches often do not have such clearly defined means of specifying their concepts.

For the beginning student, the biggest confusion is often between what is a *variable* and what is a *value* of a variable. Students will call 'male' a variable — point out to them that it is a value for a given individual, not something that *varies* across individuals.

1.1.2 Characteristics of observations

Only rarely can students define the difference between accuracy and precision. Although the thermometer is a useful introduction to the difference, the emphasis should be on how answers to questions in a survey or experimental trial can be accurate and precise, depending on the study design and the instruments used.

A good example is a question on income. If we ask for exact income in a questionnaire, we should obtain a very precise answer, possibly to the nearest cent. However, if people lie, the answers will not be very accurate. On the other hand, at the opposite extreme, if we construct the question such that people only have to indicate into which of two income groups (high or low, appropriately defined) they belong, the answers will not be precise, but lying will be much less frequent and the answers should be much more accurate than in the first case.

1.1.3 Several variables

The important point here is that, when there are several variables, they may play at least two distinct roles. Some variables simply are used to divide up a population deterministically to provide comparisons of interest whereas others have a random or probability aspect. Here, the students begin to see, for the first time, a fundamental goal of statistics: does the variability among individuals in some characteristic of interest change among different subgroups?

A Student Survey
As emphasised in the text, from the beginning, the students must become involved in the statistical approach. I find that the ideal way is to let the class choose a simple question for research about themselves, involving three binary variables, one generally being sex. This in-class project can be carried out as soon as the basic ideas about variable construction have been introduced.

Even seemingly inappropriate choices of variables, such as whether each student wears glasses or not (one does not expect a difference between the sexes) can lead to useful discussion: the time they chose this, in defining their variable, my students forgot about contact lenses. What should be done about this?

The next step after formulating the question and constructing the variables is to collect the data from the class. This should be done in two ways.

First, the $2 \times 2 \times 2$ table can be directly tabulated by going through the rows of the class and ticking for each student in the appropriate box. Especially if one question chosen involves opinion, point out that the order in which students are asked is important because earlier responses can influence the later ones. Normally, this will not occur in the usual questionnaires because they are administered individually according to some study design.

It should then be emphasised that things are not normally done by direct tabulation because there are usually many more variables, often with more than two categories. As well, a computer can do the tabulations more efficiently. Then, the students should be shown how to set up the corresponding data matrix, as in Table 1.1 of the manual, for their data in the cross-tabulated table. Generally, it is not necessary, or wise, to repeat the complete data collection process twice, especially with large classes.

1.2 Summarising data

Uses of descriptive statistics, such as graphs and frequency tables, are very important for an initial understanding of data. Only some are discussed here: frequency tables, histograms, and scattergrams, all directly related to modelling. (For example, box-plots are missing, but histograms convey more information and are central to the modelling approach.) The important thing is that they understand how to interpret them correctly.

1.2.1 Tables

The essential thing with a table is that it be clear to the reader what is being presented. Thus, labelling is critical. Important points include:

- relationships among variables will only be available by cross-classifying variables;
- proportions or percentages will usually be most appropriate for a response variable.

Indices and the sum notation should be introduced in great detail, with patience. All students can handle them if they see practically what it means that they are to do. Writing a sum out algebraically is often not enough. Examples of sums with numbers should be presented and the calculations performed by all students with a calculator.

1.2.2 Measuring size and variability

If the students are already familiar with these descriptive statistics, the weighting by n_i can cause confusion. Its role for grouped or tied data should be clearly explained.

The calculations using variances may be too advanced for some classes.

1.2.3 Graphics

Complex graphics should be avoided. As with tables, labelling is critical. Histograms and scatter plots seem to be most easy to understand intuitively. It is better that students become well acquainted with them than superficially with a wide range of graphical methods.

1.2.4 Detecting possible dependencies

The critical point here is that the student see that we are studying variability. Histograms describe one kind of variability, that for response variables. But the form of this variability may change in different segments of the population, defined by the explanatory variables.

1.3 Probability

1.3.1 Definition

Students generally have great fear of ‘probability’, often because of its reputation arising from being presented in such a dull and abstract way in terms of coins and dice. Never mention either. Talk instead about some question with binary or multiple response that you will ask to a group of people—the students in the class. When appropriate, use the answers to the three questions already collected. Then, the probabilities are just the proportions in the class who answer each way. Ideas of population and sample follow, depending on whether you are looking at everyone concerned or only a subgroup. This concrete frequency interpretation of probability generally is easily grasped.

One of the most important messages to get across near the beginning of the course is that statistics is about *groups* of individuals, and the relationships among them, and has little to say about specific individuals themselves, except as typical members of the groups. More technically, statistical models are marginal, making the assumption that differences among individuals within groups are random; for a response variable, these differences can be assumed to obey some probability distribution.

Two extremes of reactions to this idea come from sociology, where students are trained in terms of relationships among groups, and economics, based on individual rational decision-making. Obviously, the message will be more difficult to communicate in the latter case. Medicine might be intermediate, in that patients are treated individually by a doctor, but they are nevertheless classified into groups according to symptoms, and sufficiently similar patients treated identically, with varying success.

Point out to the students that, for the first two chapters, we shall ignore this distinction between sample and population. Data from groups will provisionally be analysed as if they made up the whole population.

If the criteria of exhaustiveness and mutual exclusiveness for variable construction are clearly presented, then the ideas of probability will follow naturally and easily. Make clear that, if these two criteria are not fulfilled, construction of a probability model will be impossible. It is also useful to point out that this formalisation of variable and model construction is in fact the basis of all scientific work in their discipline. In other approaches, where it is not explicitly attended to, fundamental errors can easily occur. Ways in which this can happen should be discussed.

1.3.2 Probability laws

Conditional probability is the foundation of all modelling. Thus, its definition was chosen as a basic axiom of probability. From this, important concepts such as independence follow naturally, in contrast to the product of margins definition. Use the data collected from the students to introduce this definition of conditional probability: condition on sex and compare the results to the marginal probability. The multiplicative and additive laws can then be calculated in this way, before introducing them more formally. ‘And’ events lead to multiplication; ‘or’ events to addition. The students can even see who in the table is counted twice in the additive law for nonexclusive events, if the joint probability is not subtracted.

The material on expected value may be too advanced for some classes.

1.3.3 Plotting probabilities

The important point here is that probabilities are represented by *areas* in a histogram. The concept of density is useful for calculating the sizes of the bars in a histogram. This is an occasion to introduce the concept that will be of importance later.

1.3.4 Multinomial distribution

The material on the multinomial and binomial distributions can be skipped at this stage if students are already reaching their limit with the rest of the concepts in this chapter. If it is used, build up the complete probability of observing all responses to the question obtained from the class. If the students' responses are independent (but are they, because of listening to previous answers?), the individual probabilities can be multiplied together: these are 'and' events (the probability of the first response and the second and so on). Then, this can be simplified by using powers on the probabilities. Finally, we might have asked students in some other order. All orders are mutually exclusive 'or' events because we can only use one. Without entering into details, state that the combinatorial counts the total number of possible orders in which the students could have been asked so that this also has to be included in the binomial probability.

Notice how the idea of density function can be introduced gradually, first in terms of calculating the size of the bars of a histogram (p. 30), then as a line linking the tops of the bars together (p. 33). When discussing histograms, show one with a variable having unequally spaced intervals. Illustrate how interpretation is distorted if the height, instead of the area, of each rectangle represents the probability.

The cumulative distribution function is really only introduced here in preparation for the survivor function in Chapter **??**.

1.4 Planning a study

1.4.1 Protocols

Protocols are widely used in medicine but rarely elsewhere. However, they are indispensable in any area of research.

1.4.2 Observational surveys and experiments

The basic questions of inference are raised for the first time here, only to be answered in Chapter 3.

The problem of representativity is difficult and should be discussed with the class. To illustrate it, randomly choose ten students from the class and ask them the (response) question from your student survey. The proportions giving the two answers will usually be different than for the whole class previously obtained.

The results of a statistical analysis must be convincing. The best, often the only, way to ensure that biases have not influenced the results is by randomising.

An fundamental question of study design is what is causality and how can it be studied? Students in every field must be brought to realise that, after graduation, they will be faced with these questions in virtually every study in which they are involved. For this reason, an operational definition of causality seems essential. Decision-makers will want to know what will happen if things are changed—even if the study providing the information does not involve change. Statistical honesty requires that students be aware of the limitations of observational studies.

After the demonstration that only experiments can provide direct information about causality, some students should, for at least the second time in the course, be on the point of abandoning a scientific career.

The difficulty of studying causality involving human beings and the importance of the questions being asked must be equally stressed. For me, the smoking and lung cancer question is the ideal example. Only experimentation can *directly* and unambiguously answer the question but this is impossible. Nevertheless, the question has been answered, after many observational studies, and long and acrimonious debate.

Two of the very special problems with studying human beings is that people may refuse to participate and that they may be influenced by what they know about the study. Ways of overcoming these should be discussed.

1.4.3 Study designs

Most introductory statistics courses spend little time on design of a study. Here, try to get across a few fundamental concepts, including the time orientation of the study (retrospective, cross-sectional, or prospective), whether or not there are repeated measurements (it is longitudinal or clustered), and the difference between observation and experimentation. Point out the similarity of repeated measurements to the question of dependence among answers to the class questionnaire.

1.5 Solutions to the exercises

Question (1)

Give two examples of each of the types of variables described in Section **??**, nominal, ordinal, integral, and continuous.

(a) How many possible different values does each have?

(b) For each variable, give the unit of measurement.

(c) Which may present problems in obtaining accurate results?

(d) Which do you think can be observed most precisely?

(e) For each, what will be the most appropriate way of summarising some observed data?

Answer

The answers here will depend on the choices made by the student.

Question (2)

What are the standard errors of the empirical variance and of the empirical standard deviation of a set of $n_\bullet$ observations?

Answer

This question is too difficult for this text and should not have been included.

The empirical variance is $s^2 = \sum y_i^2/n_\bullet - \bar{y}_\bullet^2$. The theoretical variance of y_i is σ_T^2 so that that of $\bar{y}_\bullet$ is $\sigma_T^2/n_\bullet$. Assume that the theoretical variance of y_i^2 is τ^2. Then, that of $\sum y_i^2/n_\bullet$ will be $\tau^2/n_\bullet$. It is reasonable to suspect that the variance of $\bar{y}_\bullet^2$ will also be $\tau^2/n_\bullet$ so that the theoretical variance of s^2 is $2\tau^2/n_\bullet$. For a normal distribution, $\tau = \sigma^2$ so that the standard error of the empirical variance can be estimated by $s^2\sqrt{2/n_\bullet}$.

It is much more difficult to show that the variance of $\sqrt{\sum(y_i - \bar{y}_\bullet)^2}$ is $\sigma_T^2/2$. Then, the standard error of the empirical standard deviation can be estimated by $s/\sqrt{2n_\bullet}$.

Question (3)

(a) Show that the theoretical variance of n_1 in the binomial distribution is equal to $n_\bullet\pi_1(1-\pi_1)$.

(b) Derive the theoretical mean $n_\bullet\pi_i$ and variance $n_\bullet\pi_i(1-\pi_i)$ of n_i for the multinomial distribution.

Answer

These questions are also difficult, but feasible for advanced students.

(a)

$$\begin{aligned}
\mathrm{E}\left[\frac{(n_1 - n_\bullet\pi_1)^2}{n_\bullet}\right] &= \sum_{n_1} \Pr(n_1, n_2)\frac{(n_1 - n_\bullet\pi_1)^2}{n_\bullet} \\
&= \sum_{n_1} \frac{n_\bullet!}{n_1!(n_\bullet - n_1)!}\pi_1^{n_1}(1-\pi_1)^{n_\bullet - n_1}\frac{n_1^2 - 2n_1 n_\bullet\pi_1 + n_\bullet^2\pi_1^2}{n_\bullet} \\
&= \sum_{n_1} \frac{n_\bullet!}{n_1!(n_\bullet - n_1)!}\pi_1^{n_1}(1-\pi_1)^{n_\bullet - n_1}\frac{n_1^2}{n_\bullet} \\
&\qquad - 2\pi_1\sum_{n_1}\frac{n_\bullet!}{n_1!(n_\bullet - n_1)!}\pi_1^{n_1}(1-\pi_1)^{n_\bullet - n_1} n_1 + n_\bullet\pi_1^2 \\
&= \sum_{n_1} \frac{(n_\bullet - 1)!}{(n_1 - 1)!(n_\bullet - n_1)!}\pi_1^{n_1}(1-\pi_1)^{n_\bullet - n_1} n_1 - n_\bullet\pi_1^2 \\
&= n_\bullet\pi_1(1-\pi_1)
\end{aligned}$$

(b)

$$\begin{aligned}
\mathrm{E}(n_i) &= \sum_{n_1}\ldots\sum_{n_I}\Pr(n_1, \ldots, n_I) n_i \\
&= \sum_{n_1}\ldots\sum_{n_I}\frac{n_\bullet!}{\prod_j n_j!}\prod_j \pi_j^{n_j} n_i
\end{aligned}$$

Table 1.1: Weights (kg) of people before and after a diet. (Dobson, 1990, p. 24)

Before	64	71	64	69	76	53	52	72	79	68
After	61	72	63	67	72	49	54	72	74	66

$$
\begin{aligned}
&= \sum_{n_1}\cdots\sum_{n_I} n_\bullet \frac{(n_\bullet-1)!}{(n_i-1)!\prod_{j\neq i} n_j!}\prod_j \pi_j^{n_j} \\
&= n_\bullet\pi_i \sum_{n_1}\cdots\sum_{n_I}\frac{(n_\bullet-1)!}{(n_i-1)!\prod_{j\neq i} n_j!}\pi_i^{n_i-1}\prod_{j\neq i}\pi_j^{n_j} \\
&= n_\bullet\pi_i
\end{aligned}
$$

That for the variance is similar to the two above.

Question (4)

In Table **??**, calculate

(a) the means and

(b) the standard deviations

before and after diet.

Answer

(a) The means are, respectively, 66.8 and 65, before and after diet.

(b) The standard deviations are, respectively, 8.42 and 7.94, before and after diet. Notice that the definition of the empirical variance in the text is the maximum likelihood estimate so that it uses division by the number of observations, not $n_\bullet - 1$. Automatic calculation using a calculator or most software will use the latter, yielding 8.88 and 8.37.

Question (5)

Calculate appropriate cross-classified percentages for the following data:

(a) the migration data of Table **??**;

(b) the car accident data of Table **??**;

(c) the myocardial infarction data of Table **??**.

In each case, discuss any relationships which may be apparent.

Table 1.2: Geographical migration among areas of Britain between 1966 and 1971. (Fingleton, 1984, p. 142)

	1971				
1966	Central Clydesdale	Lancashire & Yorkshire	West Midlands	Greater London	Total
Central Clydesdale	118	12	7	23	160
Lancashire & Yorkshire	14	2127	86	130	2357
West Midlands	8	69	2548	107	2732
Greater London	12	110	88	7712	7922
Total	152	2318	2729	7972	13171

Table 1.3: Car accidents in Florida in 1988, classified by whether or not a seat belt was worn. (Agresti, 1990, p. 30)

	Injury		
Seat belt	Fatal	Non-fatal	Total
No	1601	162527	164128
Yes	510	412368	412878
Total	2111	574895	577006

Table 1.4: Retrospective study of myocardial infarction as depending on contraceptive use. (Agresti, 1990, p. 12)

	Myocardial infarction		
Contraceptive	Yes	No	Total
Yes	23	34	57
No	35	132	167
Total	58	166	224

Table 1.5: Percentage of people in four geographical areas of Britain in 1966, given their place of residence in 1971. (from Fingleton, 1984, p. 142)

	1971				
1966	Central Clydesdale	Lancashire & Yorkshire	West Midlands	Greater London	Total
Central Clydesdale	73.8	7.5	4.4	14.4	100
Lancashire & Yorkshire	0.6	90.2	3.6	5.5	100
West Midlands	0.3	2.5	93.3	3.9	100
Greater London	0.2	1.4	1.1	97.3	100

Table 1.6: Percentages of car accidents with fatal injuries in Florida in 1988, classified by whether or not a seat belt was worn. (from Agresti, 1990, p. 30)

	Injury		
Seat belt	Fatal	Non-fatal	Total
No	0.98	99.02	100
Yes	0.12	99.88	100

Answer

(a) For the migration data of Table **??**, residence in 1971 may be expected to depend on that in 1966, so that percentages by row are appropriate. They correspond to conditional probabilities of being in a region in 1971 given one's location in 1966. These are given in Table **??**. It is evident that the vast majority of people are stable in their residence, having the same one the two years. This does not exclude their having moved within the region or having moved out and back in within the five year period. Some of the people may also have moved before or after this period. We may also note that the degree of stability increases as we move from the north to the south of the country.

(b) For the car accident data of Table **??**, the occurrence of a fatal injury may be expected to depend on whether or not a seat belt was worn, so that percentages by row are appropriate. They correspond to conditional probabilities of having a fatal injury given whether a seat belt was worn or not. These are given in Table **??**. Because of the high percentage of non-fatal accidents, these percentages might appear to be very similar. However, ratios are most informative: among those having accidents, the probability of having a fatal one is almost eight times as great without a seat belt as with one.

Care must be taken in interpreting this conclusion. We do not know why each person chose to wear a seat belt. Perhaps, careful people wear their seat belts and also have less serious accidents so that there is little direct link between wearing a seat belt and having a fatal accident. More generally, one cannot draw causal conclusions from such retrospective studies.

Table 1.7: Percentages of women with myocardial infarction as depending on contraceptive use. (from Agresti, 1990, p. 12)

	Myocardial infarction		
Contraceptive	Yes	No	Total
Yes	40.4	59.6	100
No	21.0	79.0	100

Table 1.8: Percentages of women having taken contraceptive as depending on whether they had a myocardial infarction. (from Agresti, 1990, p. 12)

	Myocardial infarction	
Contraceptive	Yes	No
Yes	39.7	20.5
No	60.3	79.5
Total	100	100

The best way to attempt to establish a causal link is to consider some design like a clinical trial where the drivers would randomly be forced to wear or not to wear a seat-belt. However obvious ethical problems arise in the present setting as there is a strong feeling that wearing a seat-belt reduces the risk of a fatal issue in a car accident (the main goal of the study probably being to quantify the risk reduction). More generally, such a study design is not usually realistic in practice, even when the 'direction' of the association between the response and one or several explanatory variables is totally unknown a priori, as one cannot, for example, assign a person to a social class.

In summary, knowledge of the study design is fundamental to ensure correct interpretation of a data analysis.

(c) For the myocardial infarction data of Table **??**, the occurrence of a myocardial infarction may be expected to depend on whether or not a contraceptive was taken, so that percentages by row would appear to be appropriate. They would correspond to conditional probabilities of having an infarction given whether a contraceptive was taken or not. These are given in Table **??**. However, the study was retrospective, with fixed numbers of women with and without an infarction, so that this table can be very misleading: it simply reflects the study design and not any new results obtained.

The only percentage table that can be legitimately calculated is that by columns and this is not too useful. It is given in Table **??** Almost twice as many of those women having an infarction had been taking a contraceptive as those without. We shall study ways of circumventing this kind of problem in the next chapter.

Question (6)

Data were collected in a study of the relationship between life stresses and illnesses. One randomly chosen member of each randomly chosen household in a sample from Oakland, California, U.S.A., was interviewed. In a list of 41 events, respondents were asked to note which had occurred within the last 18 months. The results given are for

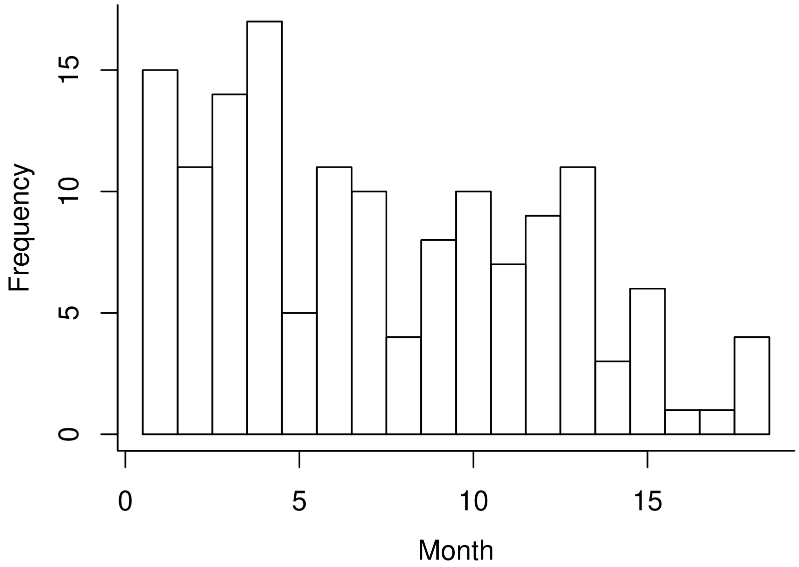

Figure 1.1: Histogram of the numbers of people having a stressful event in each of the 18 months before an interview (from Haberman, 1978, p. 3).

those recalling only one such stressful event. Our classification variable is the number of months prior to an interview that subjects remember a stressful event. Thus, the following table gives the frequency of recall of one stressful event in each of the 18 months preceding an interview (Haberman, 1978, p. 3).

Month	1	2	3	4	5	6	7	8	9
Respondents	15	11	14	17	5	11	10	4	8
Month	10	11	12	13	14	15	16	17	18
Respondents	10	7	9	11	3	6	1	1	4

Make a percentage table and a histogram of these results.

Answer

The percentages are given in Table **??** and and the corresponding histogram is shown in Figure **??**. Both show fairly clearly how the number of people recalling a stressful event is decreasing, rather irregularly, as we go back in time. Have the students discuss why both the decrease and the irregularity might have arisen.

Table 1.9: Percentage of people having a stressful event in each of the 18 months before an interview (from Haberman, 1978, p. 3).

Month	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	Total
Per cent	10.2	7.5	9.5	11.6	3.4	7.5	6.8	2.7	5.4	6.8	4.8	6.1	7.5	2.0	4.1	0.7	0.7	2.7	100.0

Question (7)

The following two tables give the observed frequencies of some (unfortunately) unspecified type of accidents (Skellam, 1948, A probability distribution derived from the binomial distribution by regarding the probability of success as variable between sets of trials. *Journal of the Royal Statistical Society* **B10**, 257–261)

Accidents	Frequency
0	447
1	132
2	42
3	21
4	3
5	2

and of car accidents in a year for 9461 Belgian drivers (Gelfand and Dalal, 1990, A note on over-dispersed exponential families. *Biometrika* **77**, 55–64, from Thyrion).

Accidents	Frequency
0	7840
1	1317
2	239
3	42
4	14
5	4
6	4
7	1

(a) Calculate the percentage tables for the two sets of frequencies.

(b) Plot the histograms and compare them.

(c) Discuss whether the first table might also refer to car accidents, keeping in mind the lapse of time between the publication of the two sets of data.

Table 1.10: Percentages of people having different numbers of an unspecified type of accident (from Skellam, 1948).

Accidents	Per cent
0	69.1
1	20.4
2	6.5
3	3.2
4	0.5
5	0.3
Total	100

Table 1.11: Percentages of Belgians having different numbers of car accidents (from Gelfand and Dalal, 1990).

Accidents	Per cent
0	82.87
1	13.92
2	2.53
3	0.44
4	0.15
5	0.04
6	0.04
7	0.01
Total	100

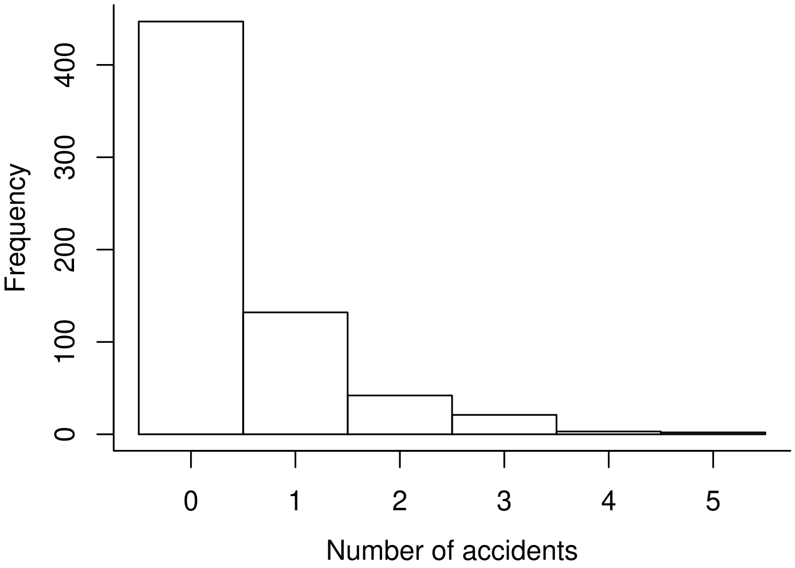

Figure 1.2: Histogram of the frequencies of an unspecified type of accident (from Skellam, 1948).

Answer

(a) The percentages are given in Tables **??** and **??**. Notice that one more decimal place is provided in the second table because of the small frequencies.

(b) The histograms are plotted in Figures **??** and **??**. We see that there are proportionally more people with no accidents in the second case. On the other hand, the presence of people with larger numbers of accidents in the second table may result simply from the larger sample size.

(c) If both tables refer to car accidents, then the level of safety has increased between the two samples, as indicated by the larger proportion with no accidents in the second case. There may also be differences between countries, because the first table is not likely from Belgium. However, there is no real evidence that the first table refers to car accidents.

Question (8)

The table below shows the numbers of units of two types of consumer goods purchased by 2000 households over 26 weeks (Chatfield, Ehrenberg, and Goodhardt, 1966, Progress on a simplified model of stationary purchasing behaviour. *Journal of the Royal Statistical Society* **B28**, 317–367; the frequency for 21 units in the last column refers to >20). The two studies were separated in time by about seven years.

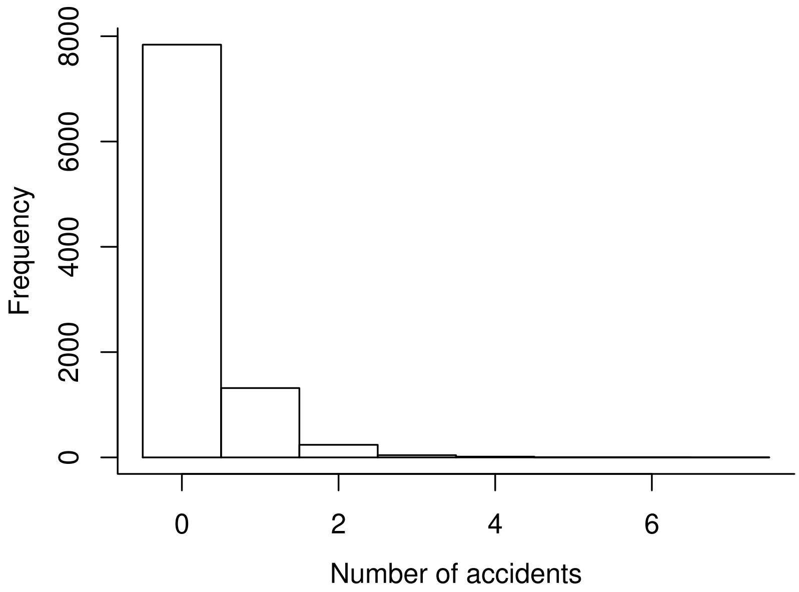

Figure 1.3: Histogram of the frequencies of Belgian car accidents in one year (from Gelfand and Dalal, 1990).

Units	Number of		Units	Number of	
bought	households buying		bought	households buying	
	Item A	Item B		Item A	Item B
0	1612	1498	14	0	2
1	164	81	15	0	2
2	71	47	16	0	3
3	47	25	17	2	1
4	28	16	18	0	0
5	17	17	19	0	2
6	12	6	20	1	1
7	12	10	21	0	12
8	5	3	22	2	
9	7	3	23	0	
10	6	6	24	0	
11	3	4	25	1	
12	3	4	26	2	
13	5	3			

(a) Why might there seem to be a somewhat larger number of people buying about 13 or 26 items?

(b) Calculate the percentage tables for the two sets of frequencies.

(c) Plot the histograms and compare them.

Answer

(a) The survey was carried out over a period of 26 weeks. Some consumers might buy the item regularly, say once a week or once fortnightly.

(b) The percentages of people buying different numbers of the items are shown in Table **??**. Notice that, in the second study, there are only 1746 households. Apparently, some were lost in the seven years between the two phases. The students should discuss how this might affect the comparison. Would it be reasonable to assume that those lost were a random subgroup from the original sample?

(c) The two histograms are given in Figures **??** and **??**. The main difference is for the purchase of zero and one items. Item A is purchased once by almost 5% more households than item B. Unfortunately, we do not know exactly what these items are. This is characteristic of the anonymity of published data from consumer surveys.

Question (9)

(a) Plot the data in Exercise (**??**) above as points on a scattergram.

(b) Does this suggest any other interpretation for these data than that from the histogram produced in the exercise above?

Table 1.12: Percentages of 2000 households buying different numbers of units of two types of consumer goods over a 26 week period (Chatfield, Ehrenberg, and Goodhardt, 1966).

Units bought	Per cent of households buying	
	Item A	Item B
0	80.60	85.80
1	8.20	4.64
2	3.55	2.69
3	2.35	1.43
4	1.40	0.92
5	0.85	0.97
6	0.60	0.34
7	0.60	0.57
8	0.25	0.17
9	0.35	0.17
10	0.30	0.34
11	0.15	0.23
12	0.15	0.23
13	0.25	0.17
14	0.00	0.11
15	0.00	0.11
16	0.00	0.17
17	0.10	0.06
18	0.00	0.00
19	0.00	0.11
20	0.05	0.06
21	0.00	0.69
22	0.10	
23	0.00	
24	0.00	
25	0.05	
26	0.10	
Total	100	100

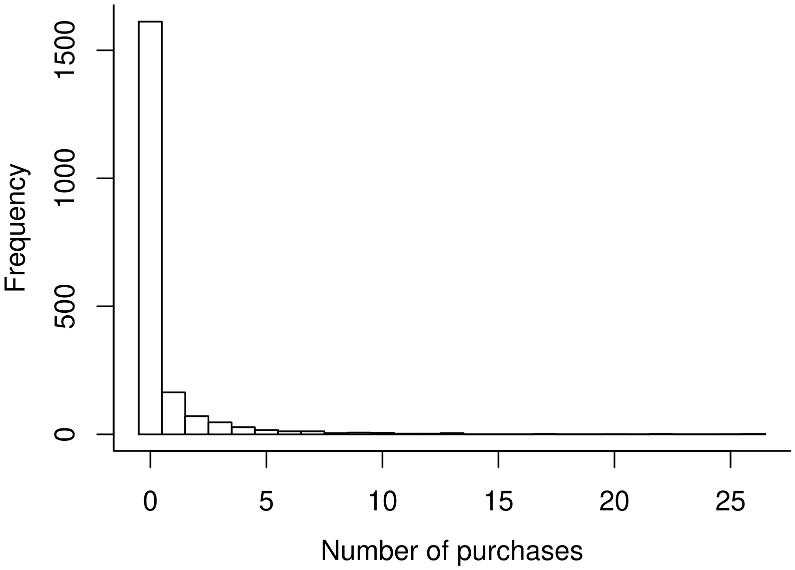

Figure 1.4: Histogram of the frequencies of purchasing different numbers of units of item A over a 26 week period (Chatfield, Ehrenberg, and Goodhardt, 1966).

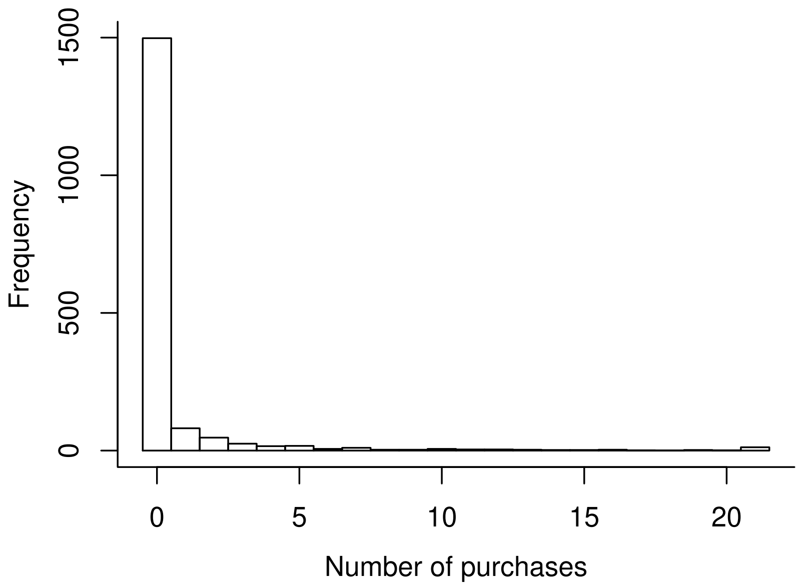

Figure 1.5: Histogram of the frequencies of purchasing different numbers of units of item B over a 26 week period (Chatfield, Ehrenberg, and Goodhardt, 1966). Note that the last category represents 21 or more purchases.

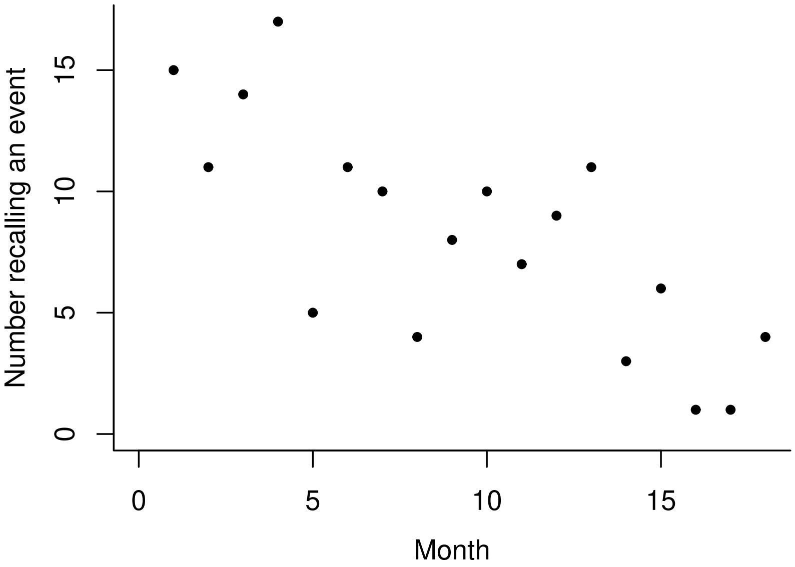

Figure 1.6: Scattergram of the numbers of people having a stressful event in each of the 18 months before an interview (from Haberman, 1978, p. 3).

Answer

(a) The scattergram of the data on stressful events is plotted in Figure **??**. We can clearly see how the number of people reporting an event decreases with time into the past. This is not just an effect of distance in the past, say forgetting. These are only the most recent events, so that all persons who give an event close in time are ineligible to give one further back in time.

(b) In spite of the great variability in the scattergram, we might think that a straight line could be traced through the points, showing how the number of events reported decreases with time into the past.

Question (10)

Consider the following two study designs.

(a) A simple random sample is drawn from women visiting a birth control clinic for the first time. They are asked whether or not they use contraceptives.

(b) A simple random sample is drawn from the list of all divorces granted in a large city over a year's time. For each couple, the length of marriage is recorded.

In each case,

(a) Describe carefully for what larger population inferences may be drawn.

(b) Give the major drawbacks of each design.

(c) Explain how you would improve the design.

Answer

(a) The birth control clinic was not chosen at random. The population could be assumed to contains all women visiting that particular clinic for the first time over some specific time period. For the divorce study, the population is all divorces in that particular city in that particular year.

(b) Neither design allows one easily to draw conclusion outside the particular clinic or city. In the first design, no information will be available about contraceptive use in the larger population. For the divorce study, no information will be available about length of marriage for people not seeking divorce. As well, the couples will have married at very different times in the past so that they may not be easy to compare.

(c) If general conclusions are to be drawn about clinics, several should be studied, drawn a random from a larger group of clinics. If conclusions are to be made about contraceptive use, information should be obtained from a population of non-clinic users, either by also including such women in the study or from other sources.

If interest centres only on that city, random sampling of cities is not necessary. However, it would be very useful to have information about lengths of marriage for a similar set of married people not divorcing in that year. The problem of differences due to dates of marriage can probably only be handled by obtained adequate explanatory variables.

Chapter 2

Categorical data

Beginning the presentation of modelling by classical simple linear regression or ANOVA has perhaps two advantages. Lecturers are very familiar with it because they were taught that way and the mathematics are simpler (if one does not present the formula for the normal distribution!). Its big drawback is that students will have no idea what it could ever be used for (which is very little in many disciplines) and what it really represents in terms of modelling (see Figure 5.1 of the manual).

Discrete data models can be related directly to the histograms and the frequency concept of probability in Chapter 1. The models describe explicitly the probabilities instead of involving some abstract parameter such as the mean of a normal distribution. As an added bonus, for the saturated models, the calculations are simpler.

2.1 Measures of dependence

2.1.1 Estimation

The first of the traditional concepts of statistics, estimation, can 'naturally' be introduced through the calculation of proportions, or relative frequencies, in a subgroup or sample. This follows directly from the presentation of probability in the first chapter. For doing the calculations of deterministic and independence relationships, replace the tables in the text by one obtained previously by 'interviewing' the students. Discuss what larger population they might reasonably represent (successive years?) in spite of the fact that they are not a random sample.

Students can easily believe that this estimation approach will provide information about the corresponding proportions, or probabilities, in the global population. Examples where a sample will not provide a reliable estimate can easily be constructed: the proportion of women visiting a birth control clinic who use contraceptives will not provide a valid estimate for the whole population of women in the region. Will the table constructed in class tell anything about students in other faculties?

A first go at introducing the important concept of degrees of freedom involves filling out a contingency table with fixed margins. Just as for the density in the first chapter, important ideas such as this should be introduced gently, in different contexts at different times, with references back to show that it is still the same thing.

2.1.2 Independence

This is a review of material from the previous chapter. The only new element is the notation. Use the results from the student questionnaire to illustrate the relationships.

2.1.3 Comparison of probabilities

The first passage through the steps of (1) differences of probabilities, (2) ratios, and (3) odds is leading up to the similar passage to logistic models. Little time need be spent on the first step here. The second, relative risk, will be of most interest to medically-oriented students who, if you criticise it, will often be prepared to argue that it is natural. Introduction of the odds ratio will bring out the sports students in the crowd.

Students should be strongly discouraged from using the term, *correlation*, to describe the relationship between variables. Point out that it is a technical term with specific meaning, only applicable in special circumstances. The (log) odds ratio performs an analogous function for *association* between binary variables.

2.1.4 Characteristics of the odds ratio

The section on the characteristics of the odds ratio is fairly technical and may not be necessary for all students. More specifically, the relationship between odds and relative risk should be especially emphasised to medical students.

2.1.5 Simpson's paradox

The manner in which Simpson's paradox is presented is very important. Basically you must show that you have nothing up your sleeves when you derive the two subtables from the global results. First, convince the students of the validity of your conclusions from the global table. Then, begin presenting the two subtables, demonstrating that they indeed add to give the original table. Draw the opposing conclusions and ask the students to explain why. This will take a bit of time because many students will want to check and recheck the calculations to verify that there is not a trick.

Emphasise that exactly the same thing can happen no matter how complex is the table, containing many variables. Introduction of a further one, that was forgotten and not even available, might drastically alter the results. On the other hand, show that each marginal table, even the two-way one is a valid average representation of the population under study (if the sampling is valid). For example, introduction of a new drug may be permitted because it will help the population on average, even although it may not be known that some people may be allergic to it. After this exercise, some students should be on the point of abandoning a scientific career.

2.2 Models for binary response variables

Finally, we get to models. Make it clear that the goal is to develop a more rigorous means of describing how histograms vary among subpopulations. Draw up simple histograms showing how they change for subgroups in the class. Later in the chapter, if the students appear to be overwhelmed by the complexity, refer back to this basic idea.

2.2.1 Models based on linear functions

Here, we go through the linear, ratio, and odds sequence for the second time. This time, emphasise the linear approach. Present the linear model as if it is the real thing. Convince them that it is the only natural way to do things: a mean probability in the population and differences from it for subgroups. Clearly show the differences, and similarity of conclusions, for the various constraints on the parameters. Then announce that this is all wrong and no one should do it this way.

If you have not been lynched, you can reassure them that, in fact, much of the technique will be used in what follows. Start into the model based on products of probabilities and show how complicated it is. Then, the students will be prepared to accept logarithms as a means of simplifying life. Of course, they should have caught on by now and will not believe that this is the correct way either. The better students will have related it all back to the sequence in the previous section.

2.2.2 Logistic models

Finally, we are ready to introduce the logistic model, which we must agree is rather complicated and does require this lead up to it. The sequence has an additional benefit of allowing the students to see that the specification of the linear part of a model, with its mean or baseline constraint, does not depend on the way this is related to the probabilities themselves.

Here it is essential that all students follow the calculations with their own calculators. Many may never have handled *log* (*ln*) and *exp* before and they need to see how they are actually calculated, even if they do not grasp the basis of them.

Work through the calculation of the parameter estimates, then calculate back to obtain the conditional probabilities, showing that they are the same as obtained by direct calculation. I have had students in the back row convinced that this was magic and would not be reproducible on another table!

An essential point with applying the logistic model to such simple tables is to show that it provides the same conclusions as can be obtained directly by inspecting the table. Emphasise that we start here with simple cases so that the students can understand exactly what is going on, but that they will soon have tables where direct understanding is not possible.

Social science students are fascinated by the sequence of tables on untouchables in India, probably partly because I can fill in details, having lived there for a year. These tables may have to be replaced by a set more appropriate to the instructor and the students. Before being asked to do a lot of calculations, the students should be given this opportunity to see results already calculated and the conclusions that can be drawn.

2.2.3 One polytomous explanatory variable

Obviously, the students will feel that 2×2 tables are not really too useful. (In many situations, they are wrong.) Luckily, the first steps in the move to larger tables are simple. The first point to get across with a polytomous explanatory variable is how to solve a system of equations. Again, the key is numerical calculation, not algebra. The second point is how the parameters describe *relationships* among the categories of the variable. The conclusions from the parameters can still be related back to the data through histograms.

Here, we can already look for simple patterns in the set of parameter estimates. Thus, we discover the contrast, from Table 2.9 in the manual, between completely free access to water (IA) and the other three categories. A second example, directly pertinent to the subject speciality of the students, might indicate a linear trend among categories. These will start the students interpreting the parameters fairly intuitively, without feeling the need to back calculate to probabilities every time.

Reasoning directly in terms of log odds can seem natural if started early enough! Emphasise that, when categorical variables have several nominal categories, no single number (like a correlation coefficient) can adequately summarise the relationships among them.

2.2.4 Several explanatory variables

Things start to get more interesting when there is more than one explanatory variable. Get the students to try to figure out which equations have to be added together to eliminate parameters. They usually will find this to be a game. Then assure them that things will really not get any more complicated than this for later models. By now, they should have gained enough confidence in their mathematical abilities.

The difficult concept here is interaction. The example for classical music is chosen especially for this, with the margin involving age showing no relationship. This is the example that should convince the students that it is all worth the trouble, that the logistic models can provide conclusions not directly visible in the table (although the histograms are pretty clear).

It is generally useful to work through the two separate conditional subtables to demonstrate that this new model with interaction gives the same results, but different for each subtable. This, and the histograms, makes the idea of interaction more concrete.

Point out that it will not make much sense to simplify the model by eliminating the parameter for age because listening habits do actually depend on age through the interaction with education. Thus, there is a hierarchy of parameters such that one should consider eliminating the more complex ones first.

Unless the class is fairly strong, the example, and exercises, with three explanatory variables may be omitted.

2.2.5 Logistic regression

Logistic regression provides the occasion to show how this type of model guarantees that the probabilities stay between zero and one. It is unfortunate that the calculations are complex. However, the algorithmic procedure (for the first iteration) generally is manageable. In strong classes, it may be interesting to explain that this is just a least squares procedure, with appropriately chosen weights. (For a general introduction to iterative weighted least squares, see Dobson, 1990, pp. 39–41.)

Here is a good opportunity to emphasise the advantage of computers. More computer-literate classes may be interested in seeing in more detail how this iterative weighted least squares procedure works. The utility of computers for repetitive tasks should be clear.

The other important point is how a model can smooth the data. The raw data on malformed children seem to indicate that a small amount of alcohol is better than none whereas the model seems to show that this could be random fluctuation. We shall have to wait until the next chapter for confirmation one way or the other.

In some classes, such as economics, now may be the time to show how dummy variables work. The Greek letters of the preceding models can be replaced by β regression coefficients multiplied by appropriately coded dummy variables, yielding models identical to those previously fitted. Thus, Equation (2.4),

$$\log\left(\frac{\pi_{1|j}}{\pi_{2|j}}\right) = \mu + \alpha_j \qquad j = 1,2$$

is identical to

$$\log\left(\frac{\pi_{1|j}}{\pi_{2|j}}\right) = \beta_0 + \beta_1 x_j \qquad j = 1,2$$

if x_j is coded (0,1) for a baseline constraint or $(-1,1)$ for a mean constraint.

2.3 Polytomous response variables

2.3.1 Polytomous logistic models

Many of the students should now be impatient to get to more 'realistic' cases where the response is not restricted two categories. Here, the detail of explanation will depend on the mathematical abilities of the class. For weak students, one can get by with the basic ideas and calculations using the algorithmic procedure. If, on the other hand, students now feel relatively at ease with manipulating logarithms and powers, a full explanation can be given. In either case, graphical presentation of changes in histograms is important.

In all cases, the algorithmic solution should be presented, because it will provide the link with multivariate analysis through log linear models below.

A series of complete analyses should be used to show how parameter values can be interpreted. The students should be instructed as to how to look for simple patterns in the matrix of α parameters. They should now feel at ease interpreting the values directly, although transformation back to ratios of probabilities is important. Here, we come back to the theme of simplifying models, again by eliminating unnecessary interactions.

2.3.2 Log linear models

Up until now, in this chapter, a single response, and a corresponding regression, have been emphasised. With log linear models, the ideas of multiple responses, and of multivariate models, are introduced. This links back to the relationships between joint and conditional probabilities in the first chapter.

The good news is that the calculations are essentially the same and that the parameter estimates are identical. However, it is important to point out that this is a unique, and valuable, property of these logistic and log linear models that will not be found in other models later in the course.

The link between the two types of models comes from the algorithmic calculations for the two-way table used above, not from some mathematical demonstration of identity of parameters. However, the demonstration is not difficult. Take the simple two-variable log linear model,

$$\log\left(\frac{\pi_{ij}}{\dot{\pi}}\right) = \mu_i + \phi_j + \alpha_{ij}$$

Suppose that the variable indexed by i is binary. Take this equation with $i = 1$ and subtract from it that for $i = 2$ to give

$$\log\left(\frac{\pi_{1j}}{\pi_{2j}}\right) = (\mu_1 - \mu_2) + (\alpha_{1j} - \alpha_{2j})$$

This can be directly related to Equation (2.4) reproduced above (if $\mu_1 = -\mu_2$ and $\alpha_{1j} = -\alpha_{2j}$, a factor of one half is involved).

In areas where retrospective studies are important, such as sociology and epidemiology, extra time should be spent on this valuable property of symmetry or reversibility of the model. No matter which variable is taken as the random response in the table, the parameters relating variables together are identical. More generally, the weight given to this material may depend on the importance of study design in your presentation. If, for example, you stress that some conditional probabilities of interest cannot be computed under retrospective designs, then careful students will already wonder why logistic and log-linear models can be estimated as if the study design was not relevant. In the context of analysing data from such a study, you could first use a logistic model and directly estimate (and interpret) the parameters in the model. Then, using the pretext that you want to check the model conclusions directly with the data, you might ask the students to compute the conditional probabilities of interest, hoping that they will remember that these are not computable with a retrospective design. You should then convince the students that the odds ratio in 'invariant' and interpretable.

This may also be an appropriate time to introduce the idea that the same types of models can be used for different kinds of prospective studies. Only the strength of conclusions will differ (causality or not) depending on whether the study was experimental (e.g. a clinical trial) or not.

Discussion of the different possible types of dependence, if it is to be presented convincingly, will require examination of subject matter tables (or perhaps the table for classical music). Again, the students should see the importance of trying to simplify a model.

2.3.3 Log linear regression

In my view, log linear (or Poisson) regression is the most important single model in all of statistics and should not be skipped. After the preceding work in this chapter, it is relatively simple and one should not expect the students to see how important it is. The calculations provide another example of weighted least squares. For very sophisticated students, it may now be possible to point out that the weights being used are just the variances for the binomial (logistic regression) and Poisson (log linear regression) distributions, although, of course, these are not actually presented until Chapter **??**.

2.3.4 Ordinal response

Ordinal models may be an optional section for many classes. However, in disciplines, such as sociology, marketing, or psychology, where such variables are central, they should be presented. The only new idea here is the reconstruction of the table, grouping categories to the left and right of various cut-points. Then, this becomes a simple application of logistic regression (although the results are approximate for the proportional odds model).

2.4 Solutions to the exercises

Question (1)

Find a study in the literature for which results are reported in the form of a contingency table.

(a) Describe the measures taken by the research workers to avoid Simpson's paradox occurring.

(b) List ways in which Simpson's paradox might have made the authors' results questionable.

(c) Invent a binary variable not used in the study and subdivide the published table in such a way that Simpson's paradox occurs.

Answer

The answers will depend on the study chosen by the student.

Table 2.1: Opinions on gun registration and the death penalty. (Agresti, 1990, p. 29)

	Death penalty		
Gun Registration	Favour	Oppose	Total
Favour	784	236	1020
Oppose	311	66	377
Total	1095	302	1397

Question (2)

Fit a logistic model to the following data:

(a) the data on injuries in car accidents and wearing seat belts of Table **??**;

(b) the data on myocardial infarction and contraceptive use of Table **??**;

(c) the data on opinions about the death penalty and gun registration of Table **??**.

In each case, discuss the meaning of the results.

Answer

(a) For the data on injuries in car accidents and wearing a seat belt of Table **??**, the logistic model is

$$\log\left(\frac{\pi_{1|j}}{\pi_{2|j}}\right) = -5.658 - 1.038x_j$$

when x_j, for wearing a seat belt, takes the values $(-1,1)$, corresponding respectively to no and yes, the mean constraint, and

$$\log\left(\frac{\pi_{1|j}}{\pi_{2|j}}\right) = -4.620 - 2.075x_j$$

when x_j takes the values $(0,1)$, the baseline constraint. In both cases, we see that the (log) odds, and the probability, of a fatal accident is considerably lower when a seat belt is worn, as indicated by the negative parameter estimate. For both models, the difference is 2.075 on the logit scale, which corresponds to an odds ratio of about one eighth $(0.125 = \mathrm{e}^{-2.075})$.

(b) For the data on myocardial infarction and contraceptive use of Table **??**, the logistic model is

$$\log\left(\frac{\pi_{1|j}}{\pi_{2|j}}\right) = -0.859 - 0.468x_j$$

when x_j takes, for contraceptive use, the values $(-1,1)$, corresponding respectively to yes and no, the mean constraint, and

$$\log\left(\frac{\pi_{1|j}}{\pi_{2|j}}\right) = -0.391 - 0.936x_j$$

Table 2.2: Percentage distribution of opinions on gun registration and the death penalty. (from Agresti, 1990, p. 29)

	Death penalty		
Gun Registration	Favour	Oppose	Total
Favour	56.1	16.9	73.0
Oppose	22.3	4.7	27.0
Total	78.4	21.6	100.0

when x_j takes the values $(0,1)$, the baseline constraint. In both cases, we see that the probability of a myocardial infarction is considerably lower when contraceptives were not taken. For both models, the difference is 0.936 on the logit scale, i.e. an odds ratio of 2.55.

This estimate provides a valid measure of the relationship between infarction and contraceptives although the study was performed retrospectively. The conditional probabilities of myocardial infarction (on contraceptive use) cannot be computed here, because the ratio of cases to controls was fixed in the study. This is reflected by the estimate of the constant (-0.859 or -0.391) in the model which is of little use. The interpretability of the model parameters contrasts with the difficulty of interpretation of the percentages in Tables **??** and **??**.

(c) For the data on opinions about the death penalty and gun registration of Table **??**, there is no clear relationship of order between the two variables whereby one would be thought to depend on the other. Hence, it may be more useful to fit a log linear model. This is

$$\log\left(\frac{\pi_{ij}}{\dot{\pi}}\right) = -0.550x_{1i} - 0.688x_{2j} - 0.087x_{1i}x_{2j}$$

when x_{1i}, for opinion on gun registration, and x_{2j}, for opinion about the death penalty, take the values $(-1,1)$, the mean constraint. The value of -0.087 ($= \hat{\gamma}_{11}$) is the same as we would have obtained from the logistic model with the same constraints. In other words, conflicting opinions appear more often together (and agreement less often) than would be expected if the two opinions were independent. This is in agreement with what one would expect from a such a study: being in favour of gun registration means being against wide use of guns and that this is associated with being against the death penalty.

These conclusions are not obvious from the percentages, as given in Table **??**. From this table, the estimated marginal probabilities are $\hat{\pi}_{1\bullet} = 0.730$ and $\hat{\pi}_{\bullet 1} = 0.784$ of having favourable opinions on the two questions. Multiplying them together, we obtain $\tilde{\pi}_{11} = 0.57$ as the estimated probability of having both opinions favourable, if answers to the two questions were independent. The observed value, $\hat{\pi}_{11} = 0.56$, is slightly less, confirming our analysis with the model.

Some care must be taken with judgements as to whether a parameter (such as the association parameter γ_{11} above) is small or not. Several questions should be asked:

- Is the parameter really different from zero or is the estimate only randomly different, a result of sampling? This will be covered in Chapter **??**.

- If the parameter is indeed different from zero,
  - is the estimate relatively smaller than estimates of other similar parameters in the model?
  - is the estimate large enough to have any substantive meaning in the field of study?

Of course, small values can be very important, if a large value was expected. So-called 'negative' results, where a parameter can be eliminated from a model (because it is non-significant), may lead to scientific discoveries as easily as large estimates.

Question (3)

(a) Fit a log linear model to the migration data of Table **??**.

(b) Explain how residence in 1971 is related to that in 1966.

(c) Notice that the four areas are ordered from the north to the south of Britain. Can a better model be constructed using this information?

Answer

(a) For a log linear model with mean constraints, the parameter estimates for these data have $(-1.378, 0.443, 0.245, 0.690)$ for the 1966 margin, $(-1.469, 0.308, 0.218, 0.943)$ for the 1971 margin, and

$$\begin{pmatrix} 3.156 & -0.907 & -1.356 & -0.893 \\ -0.798 & 2.449 & -0.669 & -0.982 \\ -1.159 & -0.781 & 2.918 & -0.978 \\ -1.199 & -0.761 & -0.893 & 2.853 \end{pmatrix}$$

for the α parameters relating the two dates together.

(b) It is clear that there is a very high probability of a person being in the same region on the two dates. Most of the off-diagonal values are fairly similar, close to -1.0. (Moves from Central Clydesdale to the West Midlands are relatively somewhat less frequent and those from Lancashire and Yorkshire to the West Midlands slightly more frequent.) There is no indication that the probabilities decrease in value with distance between regions, i.e. with distance from the main diagonal. Thus, the estimated probability of moving between any pair of regions is about constant.

(c) Because the regions are ordered, we might want to use a model for ordinal data. However, the ones that we have studied only apply to an ordinal response, here the place of residence in 1971, and would not take into account the symmetry between the two dates.

Question (4)

The following table (Fienberg, 1977, p. 16) gives data on the choice of piano by soloists playing for selected major American orchestras during the 1973–1974 concert season in the U.S.A.

	Piano	
Orchestra	Steinway	Other
Boston	4	2
Chicago	13	1
Cleveland	11	2
Minnesota	2	2
New York	9	2
Philadelphia	6	0

(a) Study the relationship between the two variables.

(b) Might a log linear model be more appropriate than a logistic model?

(c) The same soloist may have appeared with different orchestras. Discuss what difficulties this may create for the models which you have used.

Answer

(a) Because of the zero in the table, we shall approximate it by 1/2 in the calculation of the parameter estimates. With the mean constraint in a logistic model, we obtain $\hat{\mu} = 1.492$, $\hat{\alpha}_1 = -0.799$, $\hat{\alpha}_2 = 1.073$, $\hat{\alpha}_3 = 0.213$, $\hat{\alpha}_4 = -1.492$, $\hat{\alpha}_5 = 0.012$, and $\hat{\alpha}_6 = 0.993$. As is evident from the table (without the approximation), Chicago and Philadelphia ($\hat{\pi}_1 = 1$) have the highest probabilities of using a Steinway and Minnesota the lowest (although the latter is still estimated as 0.5).

(b) The choice of piano might not be thought really to depend on the orchestra (except indirectly through choice of soloists) so that we would then be interested simply in the association between type of piano and orchestra. In this case, a log linear model would be more appropriate, but of course the results will be identical.

(c) One of the hypotheses behind the use of logistic and log linear models is that the frequencies in a table are composed of independent events. If the same soloist played several times, whether for the same or different orchestras, he or she would probably use the same type of piano and the events would not be independent.

Question (5)

The table below gives the frequency of coronary heart disease by age group (Hosmer and Lemeshow, 1989, p. 4). The latter was originally measured in years, but larger groupings were then created.

	Coronary heart disease	
Age	Yes	No
20–29	9	1
30–34	13	2
35–39	9	3
40–44	10	5
45–49	7	6
50–54	3	5
55–59	4	13
60–69	2	8

(a) Fit a logistic regression model to these data.

(b) Plot and interpret the results.

(c) What difficulties would you have encountered in making the plot if you had used the original raw data with the actual ages, in years, of the 100 people involved?

Answer

(a) The age classification $(20,29),(30,34),\ldots,(60,69)$ given in this exercise stands for the age intervals $(19.5,29.5),(29.5,34.5),\ldots,(59.5,\ 69.5)$. Hence the interval midpoints for the ages are $(24.5,32,37,\ 42,47,52,\ 57,64.5)$. Note that these values are simply the average of the originally given bounds. There are situations where this rule does not apply: these will be mentioned when appropriate.

The approximate estimates for the logistic regression are $\hat{\beta}_0 \doteq 4.984$ and $\hat{\beta}_1 \doteq -0.104$ and the exact maximum likelihood estimates are $\hat{\beta}_0 = 5.038$ and $\hat{\beta}_1 = -0.105$. The negative slope indicates that the probability of coronary heart disease is decreasing with age.

(b) The estimated proportions of people having coronary heart disease and the fitted logistic regression line are plotted in Figure **??**. As expected, we see how the probability of heart disease decreases with age. The proportion with the disease decreases from about 90% at age 25 to about 20% at age 60. The fitted curve is reasonably close to the observed proportions. This is obviously not a random sample from an ordinary population.

(c) With the raw data, there would likely be few people at any given age so that the estimated proportions would often be zero or one and would jump rather erratically between these two values. However, the estimation of the logistic regression curve would be more accurate if the exact ages could be used.

As a general comment, it would be interesting to know how this sample was chosen. It seems doubtful that this model could represent some typical population.

Question (6)

In Section 2.2.4, we studied data on listeners to classical music radio programmes (Table **??**). The table below gives similar data on listening to religious and to discussion programmes on the radio (Lazarsfeld, 1955).

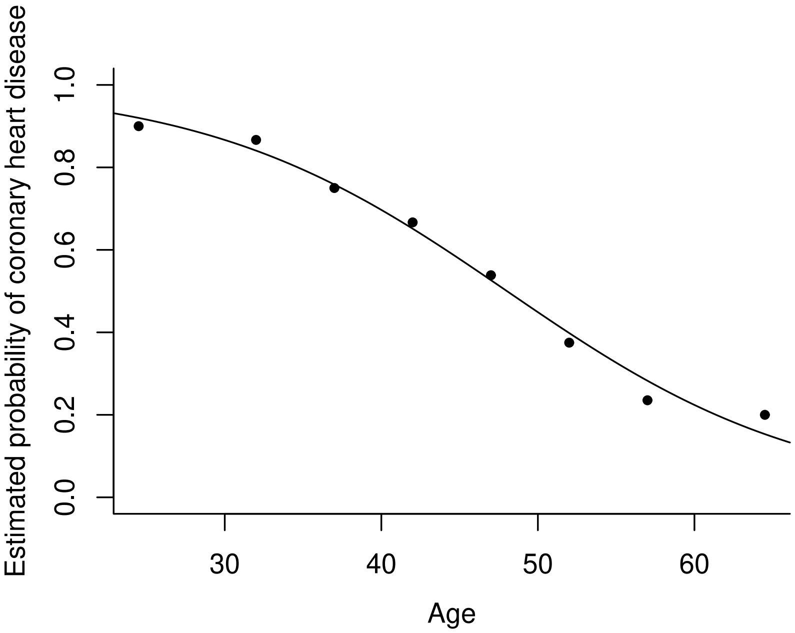

Figure 2.1: Scattergram of the estimated proportions of people having coronary heart disease at different ages (from Hosmer and Lemeshow, 1989), with the fitted logistic regression line.

Table 2.3: Classification of classical music listeners by age and education. (Lazarsfeld, 1955)

	Education			
	High		Low	
	Listen to classical music			
Age	Yes	No	Yes	No
Old	210	190	170	730
Young	194	406	110	290

	Education							
	High		Low		High		Low	
	Listen to							
	religious programmes				discussion programmes			
Age	Yes	No	Yes	No	Yes	No	Yes	No
Old	45	355	285	615	210	180	360	540
Young	55	545	115	285	240	360	100	300

The definitions of the age and education variables are the same as stated in the text above. The rounded values in the tables result because of the stylised nature of the data, already mentioned above in the text.

(a) Check that the joint distribution of age and education is the same in all three tables.

(b) Fit an appropriate logistic model to each half of the table (each type of programme).

(c) Are the results similar to those given above for classical music?

(d) Can you explain why?

Answer

(a) For all three types of programmes, the marginal table for age and education is

	Education	
Age	High	Low
Old	400	900
Young	600	400

showing that the joint distribution is the same in all three cases. This would be expected if all three tables arise from the same study (unless there were missing values in some of the response variables).

(b) We shall use the same definitions of the variables as for classical music in the text: age indexed by j (α_j) and education by k (β_k). For religious programmes, the parameter estimates are $\hat{\mu} = -1.509$, $\hat{\alpha}_1 = 0.092 = -\hat{\alpha}_2$, $\hat{\beta}_1 = -0.671 = -\hat{\beta}_2$, and $\hat{\gamma}_{11} = 0.022 = -\hat{\gamma}_{12} = -\hat{\gamma}_{21} = \hat{\gamma}_{22}$. For discussion programmes, they are $\hat{\mu} = -0.439$, $\hat{\alpha}_1 = 0.313 = -\hat{\alpha}_2$, $\hat{\beta}_1 = 0.313 = -\hat{\beta}_2$, and $\hat{\gamma}_{11} = -0.033 = -\hat{\gamma}_{12} = -\hat{\gamma}_{21} = \hat{\gamma}_{22}$.

(c) The results are quite different in the three cases. In the two just studied, there is little evidence of interaction between age and education. For religious programmes, there is little difference with age, whereas more highly educated people listen less. For discussion programmes, both older and more highly educated people listen more, the effect being cumulative.

(d) Thus, there seems to be a change in listening habits with age, in respect to education level, for classical music but not for the other two types of programmes. But the age effect could also arise from a difference in education when the two age groups were young.

Question (7)

Consider data on the relationship among delinquency, socioeconomic status (SES), and being a boy scout, given below (Agresti, 1990, p. 157).

		Delinquent	
SES	Scout	Yes	No
Low	Yes	10	40
Low	No	40	160
Medium	Yes	18	132
Medium	No	18	132
High	Yes	8	192
High	No	2	48

(a) Fit a logistic model to explore the relationships among the variables.

(b) What is peculiar about these data? Look at models for the marginal tables, when delinquency depends on only one of the explanatory variables.

(c) Relate your conclusions to the difficulties in drawing causal conclusions from sample survey data.

Answer

(a) For the logistic model, the parameter estimates are $\hat{\mu} = -2.186$, $\hat{\alpha}_1 = 0.000 = -\hat{\alpha}_2$, $\hat{\beta}_1 = 0.799$, $\hat{\beta}_2 = 0.193$, $\hat{\beta}_3 = -0.992$, $\hat{\gamma}_{11} = 0.000 = -\hat{\gamma}_{21}$, $\hat{\gamma}_{12} = 0.000 = \hat{\gamma}_{22}$, and $\hat{\gamma}_{13} = 0.000 = \hat{\gamma}_{23}$, where j indexes being a scout or not and k indexes SES. Delinquency depends only on SES and not on whether the boy was a scout or not. The probability of being a delinquent decreases in the higher SES.

(b) We might already suspect something strange with this table because the two rows for medium SES are identical. The marginal table for delinquency and scout is

	Delinquent	
Scout	Yes	No
Yes	36	364
No	60	340

and the logistic model has estimates, $\hat{\mu} = -2.024$ and $\hat{\alpha}_1 = -0.290 = -\hat{\alpha}_2$. When the effect of SES is ignored, we find that delinquency depends on being a scout.

The marginal table for delinquency and SES is

	Delinquent	
SES	Yes	No
Low	50	200
Medium	36	264
High	10	240

with estimates, $\hat{\mu} = -2.186$, $\hat{\beta}_1 = 0.799$, $\hat{\beta}_2 = 0.193$, and $\hat{\beta}_3 = -0.992$. These are identical to those in the full model above.

If we look back at the original table, we see that the proportion of delinquent boys is exactly the same, within each level of SES, whether the boys are scouts or not. Thus, participation in the scouts is associated with SES, explaining the misleading relationship. This relationship can be studied by constructing the appropriate marginal table.

(c) If we only had the variable, being a scout, available, we would have concluded that the probability of delinquency depended on this variable. However, the introduction of SES shows this not to be the case.

In sample survey data, some crucial explanatory variable may always be missing, making any 'causal' conclusions misleading. This is a form of Simpson's paradox. Conclusions can change drastically when a key explanatory variable is introduced.

Question (8)

In Section 2.2.4, we looked at a study of knowledge of cancer, given in Table **??**. The table below reproduces the data, but with the 'lectures' variable replaced by 'serious reading' (Lombard and Doering, 1947).

		Radio			
		Yes		No	
		Knowledge			
Newspaper	Reading	Good	Poor	Good	Poor
Yes	Yes	125	75	228	195
	No	43	63	82	162
No	Yes	17	19	70	91
	No	17	53	86	403

(a) Fit a logistic model and interpret the results.

(b) Compare them with those given above and explain any differences.

(c) If you have appropriate computer software available, look at models for all four explanatory variables simultaneously.

Table 2.4: Sources of knowledge of cancer. (Lombard and Doering, 1947)

		Radio			
		Yes		No	
		Knowledge			
Newspaper	Lectures	Good	Poor	Good	Poor
Yes	Yes	31	12	34	24
	No	137	126	276	333
No	Yes	5	6	5	18
	No	29	66	151	476

			Reading			
			Yes		No	
			Knowledge			
Radio	Newspaper	Lectures	Good	Poor	Good	Poor
Yes	Yes	Yes	23	8	8	4
Yes	Yes	No	102	67	35	59
Yes	No	Yes	1	3	4	3
Yes	No	No	16	16	13	50
No	Yes	Yes	27	18	7	6
No	Yes	No	201	177	75	156
No	No	Yes	3	8	2	10
No	No	No	67	83	84	393

(d) Again, compare the results with those from the simpler tables and explain any differences.

Answer

(a) Let j index radio, k reading, and l newspapers. The parameter estimates are $\hat{\mu} = -0.431$, $\hat{\alpha}_1 = 0.152$, $\hat{\beta}_1 = 0.505$, $\hat{\delta}_1 = 0.332$, $\hat{\gamma}_{111} = -0.025$, $\hat{\gamma}_{211} = 0.012$, $\hat{\gamma}_{311} = -0.072$, and $\hat{\gamma}_{4111} = 0.039$. Here, good knowledge of cancer is most strongly positively associated with reading. This is one and a half times stronger than that for newspapers, which is, in turn, about twice as strong as that for radio. There is little indication of interactions among the sources of knowledge.

(b) The parameters for dependence of cancer knowledge on radio and newspapers are about one half the size in the previous model in the text. However, that for newspapers is still about twice as large as for radio. In that model, lectures were most weakly linked with knowledge. The variable that replaces it, reading, is the most strongly associated. Apparently, people who obtain knowledge from reading also do so from radio and newspapers, explaining the reduction in the latter relationships when the former is introduced into the model.

(c) Let j index reading, k lectures, l newspapers, and m radio. The parameter estimates are $\hat{\mu} = -0.306$, $\hat{\alpha}_1 = 0.243$, $\hat{\beta}_1 = 0.271$, $\hat{\delta}_1 = 0.507$, $\hat{\tau}_1 = 0.170$, $\hat{\gamma}_{111} = -0.113$, $\hat{\gamma}_{211} = -0.031$, $\hat{\gamma}_{311} = 0.031$, $\hat{\gamma}_{411} = 0.128$, $\hat{\gamma}_{511} = -0.289$, $\hat{\gamma}_{611} = 0.207$, $\hat{\lambda}_{1111} = 0.137$, $\hat{\lambda}_{2111} = -0.125$, $\hat{\lambda}_{3111} = -0.043$, $\hat{\lambda}_{4111} = 0.140$, and $\hat{\lambda}_{51111} = 0.129$. Now, good knowledge of cancer is most strongly positively associated with newspapers. Lectures are now second with reading a close third. However, there are many large interactions which make interpretations very difficult. (Fortunately, many of them are unnecessary. After simplifying the model using the AIC, as in the next chapter, those left are between newspapers and lectures, newspapers and reading, and lectures and reading.)

(d) The existence of so many interactions shows that any interpretations of the smaller tables can be misleading.

Question (9)

A study was conducted to determine factors which might influence shopping behaviour. The sample was taken at random from the population of the town of Dukinfield, Greater

Manchester, England. In the following table, we present the variables, choice of shopping centre, age, income, and car ownership (Fingleton, 1984, p. 25).

		Car owner			
		Yes		No	
		Shopping centre			
Age	Income	Near	Other	Near	Other
Young	Low	12	57	17	48
	High	3	24	2	3
Old	Low	18	53	51	105
	High	2	11	1	0

Unfortunately, the author does not state how the categories for the variables were constructed.

(a) Fit a logistic model.

(b) Interpret the results.

Answer

(a) The shopping behaviour under study is the distance to the shopping centre and the factors that might determine this, car ownership, income, and age. To calculate the parameters, we approximate the (sampling) zero in the table by 1/2. Let j index car ownership, k income, and l age. Then, the parameter estimates are $\hat{\mu} = -0.987$, $\hat{\alpha}_1 = -0.619$, $\hat{\beta}_1 = -0.113$, $\hat{\delta}_1 = -0.283$, $\hat{\gamma}_{111} = 0.399$, $\hat{\gamma}_{211} = 0.070$, $\hat{\gamma}_{311} = 0.085$, and $\hat{\gamma}_{4111} = -0.111$.

(b) All of the interactions, except that between car ownership and income, are reasonably small, as is the estimate for income itself. The younger age group tends to shop further away. However, car ownership gives the strongest relationship. On average car owners shop further away (-0.619). This is much less true of the low income (-0.220) than the high income (-1.018) group.

Question (10)

The following table shows the numbers of household burglaries in Detroit, U.S.A., 1974–1975, obtained from the National Crime Survey (Nelson, 1980).

Number of burglaries	Number of households
0	8385
1	976
2	183
3	35
4	5
5	2

(a) Fit a log linear regression model to these data.

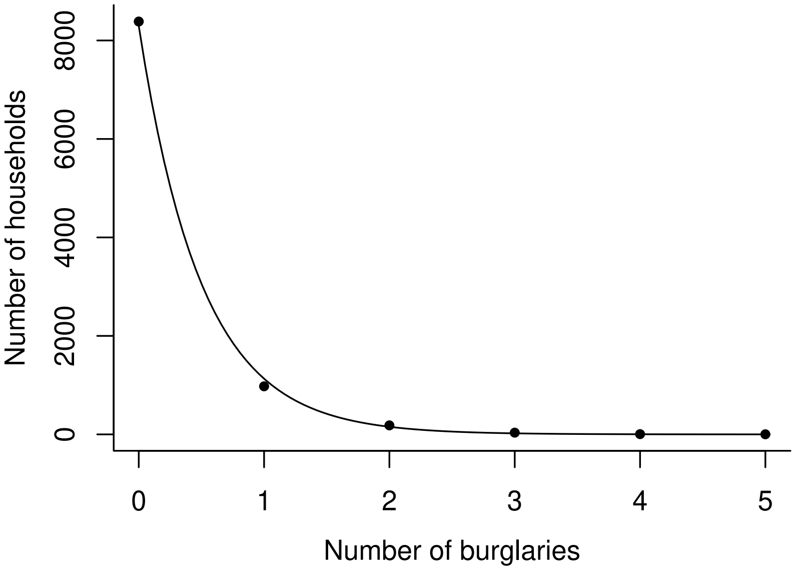

Figure 2.2: Scattergram of the estimated numbers of households having burglaries (Nelson, 1980), with the fitted log linear regression line.

(b) Plot and interpret the results.

Answer

(a) The goal here is to see if we can find a simple smooth relationship between the number of burglaries and their frequency. (We shall study more reasonable ways in Chapter **??**.) The approximate estimates for the log linear regression are $\hat{\beta}_0 \doteq 9.024$ and $\hat{\beta}_1 \doteq -1.993$. (The exact maximum likelihood estimates are $\hat{\beta}_0 = 9.025$ and $\hat{\beta}_1 = -2.013$.) The negative slope indicates that the probability decreases as the number of burglaries increases.

(b) The estimated frequencies of households having burglaries and the fitted log linear regression line are plotted in Figure **??**. The model fits the data very closely. Households with several burglaries are much rarer than those with few. Even the zero category fits well, showing that they are not a special group.

Question (11)

The table in Exercise (**??**) gave the frequency of recall of a stressful event over an 18 month period.

(a) Study how recall depends on time by fitting a log linear regression model.

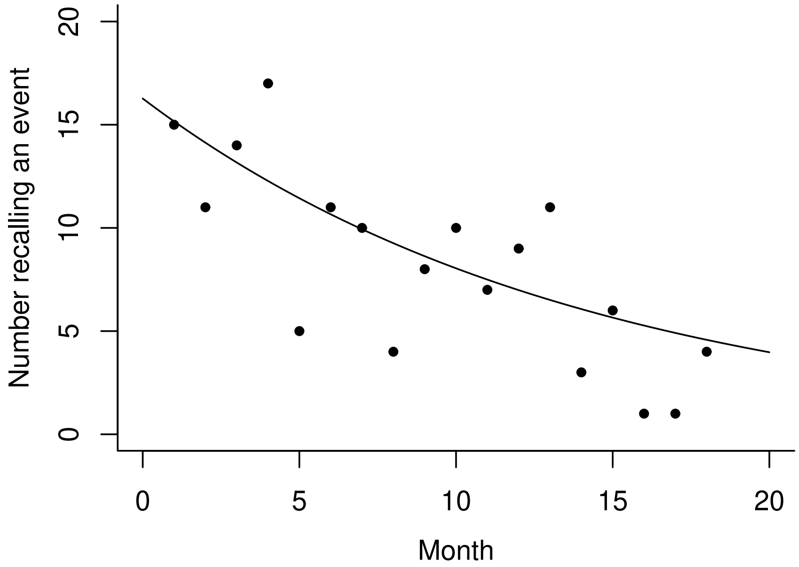

Figure 2.3: Scattergram of the estimated numbers of events recalled each month in the past (from Haberman, 1978, p. 3), with the fitted log linear regression line.

(b) Plot and interpret the results.

Answer

(a) The approximate estimates for the log linear regression are $\hat{\beta}_0 \doteq 2.789$ and $\hat{\beta}_1 \doteq -0.070$. (The exact maximum likelihood estimates are $\hat{\beta}_0 = 2.803$ and $\hat{\beta}_1 = -0.084$.) The negative slope indicates that the frequency of events remembered decreases with the number of months in the past.

(b) The estimated frequencies of recall of stressful events over time and the fitted log linear regression line are plotted in Figure **??**. These data are rather scattered so that the fitted line only goes more or less through the middle of them. As we saw and discussed in Exercise (**??**), the frequency of events decreases with time into the past. The log linear model describes this well.

Question (12)

People involved in a driver education study were followed over a four-year period. Traffic violations each year among male subjects in the control group were recorded as shown in the following table (Davis, 2002, p. 228).

Year				
1	2	3	4	
No	No	No	No	731
No	No	No	Yes	310
No	No	Yes	No	256
No	No	Yes	Yes	196
No	Yes	No	No	156
No	Yes	No	Yes	121
No	Yes	Yes	No	114
No	Yes	Yes	Yes	152
Yes	No	No	No	61
Yes	No	No	Yes	40
Yes	No	Yes	No	45
Yes	No	Yes	Yes	39
Yes	Yes	No	No	47
Yes	Yes	No	Yes	42
Yes	Yes	Yes	No	46
Yes	Yes	Yes	Yes	53

(a) Develop a log linear model to describe the association between violations in the various years.

(b) Is the association stronger for years closer together in time?

(c) Is it reasonable to simplify the model by only including associations between adjacent years?

Answer

(a) After eliminating unnecessary interactions, the following relationships remain: between years 1 and 2: 0.236; between years 1 and 3: 0.100; between years 2 and 3: 0.118; between years 2 and 4: 0.138; between years 3 and 4: 0.134; among years 1, 2, and 3: 0.063.

(b) The strongest association is between years 1 and 2. That between years 1 and 4 is unnecessary. That between 2 and 4 is somewhat larger than those between 2 and 3 or 3 and 4.

(c) No, it is not possible to simplify the model in this way. There are long term dependencies.

Question (13)

The following table gives the results of a social survey of income and job satisfaction in the U.S.A. (Agresti, 1990, p. 21). They are taken from the 1984 General Social Survey of the National Data Program.

	Satisfaction			
	Very	Little	Moderately	Very
Income	dissatisfied	dissatisfied	satisfied	satisfied
< \$6000	20	24	80	82
\$6000–14999	22	38	104	125
\$15000–24999	13	28	81	113
≥ \$25000	7	18	54	92

(a) Fit a polytomous logistic model to these data and interpret the results.

(b) What might be a more appropriate model?

Answer

(a) For a log linear model with mean constraints, the parameter estimates for these data have $(0.020, 0.330, 0.034, -0.384)$ for the income margin, $(-1.071, -0.461, 0.632, 0.900)$ for the satisfaction margin, and

$$\begin{pmatrix} 0.326 & -0.102 & 0.010 & -0.234 \\ 0.112 & 0.048 & -0.038 & -0.122 \\ -0.119 & 0.039 & 0.008 & 0.072 \\ -0.319 & 0.015 & 0.020 & 0.284 \end{pmatrix}$$

for the α parameters relating the two together. Only the four extreme values of the associate matrix are large. The concordant corners, for example low income and very dissatisfied, have higher probability and the discordant ones, such as high income and very dissatisfied, have lower probability.

(b) Both variables are ordered. The models for ordinal variables might be used but they only handle an ordinal response variable. (In fact, for these data, a linear interaction fits very well and has eight fewer parameters than the model fitted above.)

Question (14)

The table below shows party affiliation and political ideology of a sample of voters during the 1976 Wisconsin, U.S.A., presidential primary election (Agresti, 1984, p. 87).

	Political ideology		
Party	Liberal	Moderate	Conservative
Democrat	143	156	100
Independent	119	210	141
Republican	15	72	127

(a) Fit the continuation ratio and the (approximate) proportional odds models to these data.

(b) Compare and interpret the results.

Answer

(a) The reconstructed table for the continuation ratio models is

	Ideology	
Party	Left	Right
Democrat	143	156
Independent	119	210
Republican	15	72
Democrat	299	100
Independent	329	141
Republican	87	127

The parameters of interest in the logistic model are $\hat{\alpha} = (0.614, 0.250, -0.864)$.

The reconstructed table for the proportional odds model is

	Ideology	
Party	Left	Right
Democrat	143	256
Independent	119	351
Republican	15	199
Democrat	299	100
Independent	329	141
Republican	87	127

The parameters of interest in the approximate logistic model are $\hat{\alpha} = (0.704, 0.330, -1.034)$.

(b) Democrats have a higher probability of being to the left of the political spectrum. Independents are also somewhat left of centre whereas Republicans have a high probability of being to the right. The results are for the two models are similar but the second is somewhat more extreme. The interaction parameters, which should be close to zero if the model is acceptable, are $\hat{\gamma} = (0.040, -0.076, 0.036)$ and $\hat{\gamma} = (0.130, 0.004, -0.135)$ respectively. The latter are larger, indicating that the former model, the continuation ratio, may be preferable.

Question (15)

The following table gives ratings of the performance of radio and television, by the person's colour for samples taken in two different years (Agresti, 1984, p. 103).

		Rating		
Year	Colour	Poor	Fair	Good
1959	White	54	253	325
	Black	4	23	81
1971	White	158	636	600
	Black	24	144	224

(a) Fit the continuation ratio and the (approximate) proportional odds models to these data.

(b) Compare and interpret the results.

Answer

(a) The reconstructed table for the continuation ratio models is

		Rating	
Year	Colour	Poor	Good
1959	White	54	253
	Black	4	23
1971	White	158	600
	Black	24	224
1959	White	307	325
	Black	27	81
1971	White	794	600
	Black	168	224

The parameters of interest in the logistic model are $\hat{\alpha}_1 = 0.277$ for colour, $\hat{\beta}_1 = -0.157$ for year, and $\hat{\gamma}_{11} = 0.035$ for the interaction between the two.

The reconstructed table for the proportion odds model is

		Rating	
Year	Colour	Poor	Good
1959	White	54	578
	Black	4	104
1971	White	158	1236
	Black	24	368
1959	White	307	325
	Black	27	81
1971	White	794	600
	Black	168	224

The parameters of interest in the approximate logistic model are $\hat{\alpha}_1 = 0.396$ for colour, $\hat{\beta}_1 = -0.249$ for year, and $\hat{\gamma}_{11} = 0.086$ for the interaction between the two.

(b) White people tend to give a poorer rating of the performance of radio and television. There is some indication that ratings have improved between the two years. The interaction is small indicating that it has improved in about the same way for both blacks and whites. As in the previous exercise, the latter model gives larger estimates than the former. The interactions of these variables with the two subtables are $(-0.126, 0.130, -0.084)$ for the continuation ratio whereas they are $(-0.006, 0.038, -0.032)$ for the proportional odds. In this case, the latter model seems to fit slightly better.

Chapter 3

Inference

In this chapter, we finally reach what many statisticians believe to be the core of statistical thinking: inference. We now must face the problem of how to draw conclusions in the presence of random variability of the observations in a sample drawn from a population. When the likelihood function is used, this becomes a relatively trivial problem.

Students in traditional introductory statistics courses are fed a stream of hypothesis testing procedures, based on the Chi-squared, Student t, and F distributions. When they face a problem in the real world, they can never figure out which one should be applied. (I recall vividly being in that situation during my first job.) The unlucky ones cannot even remember if it is the Chi-squared value or the P-value that is supposed to be small!

Many statisticians seem to think that likelihood is a complex and difficult concept. For new, unindoctrinated, students, likelihood can be simple and intuitive, whereas testing is confusing and contradictory. No wonder that statistics has such a bad name.

3.1 Goals of inference

In fact, the basic inference problems have already appeared in the presentations of probability and models in the previous chapters. When a model smooths the observed data, how do we know if the differences between the model probabilities and the observed relative frequencies are random or indicative of missed structure in the data? How can we judge if a model can be simplified, say by eliminating an interaction?

3.1.1 Discovery and decisions

The orientation of the course will depend on the type of students. Students heading for research will require the emphasis in the book. Those in engineering and perhaps economics and finance, will require primarily hypothesis testing. Students in medicine should be familiar both with model selection based on the likelihood and with testing.

3.1.2 Types of model selection

The basic point to transmit is that, in scientific research, all of these model selection problems are closely related aspects of the same problem. This is completely hidden in classical statistics.

3.2 Likelihood

3.2.1 Likelihood function

The necessity for inference procedures must proceed from the demonstration that probabilities calculated from a sample will not necessarily be identical to those in the global population. Simple examples of questionnaires, with binary response to a question, usually suffice. Thus, draw a random sample of ten students from the class and ask again the question from the survey of Chapter **??**. (I already suggested this for Section **??**, but it will not hurt to do it again.) Discuss the chances of getting the same result as in the original survey. Then, make the exact probability calculations (results are in Table 3.1!

Then, the question arises as to what we can do with the fixed information in the sample that we have obtained. The contrast must be made between a clearly specified model, before data collection, that allows us to predict what might happen, and fixed observations after, that allow us to infer what model might have produced them. Reading Table 3.1 in the two directions should get this across. In this binomial case, contrast the difference between the small number of different possible observations with the infinite number of different models.

The definition of likelihood follows intuitively. It should be repeated slowly several times so that the students can think about it. Stress that this is a fundamental key to the basic philosophical question of trying to draw general conclusions from specific observations.

3.2.2 Maximum likelihood estimate

The idea that the likelihood function has a maximum will generally be presented empirically, from Table 3.1, or other examples. Few non-mathematics students can be expected to understand derivatives, so that this should be omitted.

3.2.3 Normed likelihood and deviance

Several important points need to be emphasised. Likelihood only allows a relative comparison of models, hence the use of norming. Once again, logarithms can simplify life by making things additive, yielding the deviance. The likelihood function can be plotted, but often summaries, such as some sort of likelihood interval, are sufficient. As the sample size increases, we would expect to have more information available, i.e. the parameters of interest will be known more precisely, and indeed the likelihood function does become narrower.

Students must clearly understand that deviance provides a metric for measuring the distance between models. They must see that they have to know what two models are being compared in any given instance. Then the larger is the deviance, the further apart are the two models. Because the more complex model will be closer to the data, a large deviance will be an indication that the simpler model is unacceptable—it makes the data too much less probable than the more complex model.

We can now refer back to one of our problems of the previous chapter. When can a parameter, such as an interaction, reasonably be set to zero? Here, it must be demonstrated that this does not really depend on how small the parameter is, but rather on the width of the likelihood function, i.e. on how uncertain, from the data, we are about its value, how much information we have about it (Figure 3.2 in the text).

3.2.4 Standard errors

Approximations have been important in statistics, although their role is finally decreasing with the growing power of computers. However, the standard error really cannot be skipped. Such a difficult concept to explain! It really must be taken on faith that the normal approximation in the middle of p. 123 will approximate a likelihood function. The dangers of this approximation, especially with small samples, should be stressed. Intervals obtained from the normal approximation to the likelihood, based on standard errors, can contain impossible values, such as probabilities that are negative or greater than one!

3.3 Two special models

3.3.1 Saturated models

Although a saturated model of some sort will always exist, their primary use is for logistic and log linear models with contingency tables. There, a unique saturated model exists, with deviance 0.

3.3.2 Null models

We can now finally address a more realistic question of model simplification: can certain parameters in a logistic or log linear model reasonably be set to zero? Because the problem is multidimensional, the explanation is really only feasible in terms of the underlying probabilities and the question of independence. Students should be walked through the calculations involved in the fundamental formula of Equation (3.6).

Note that it does not make sense to set just any parameter to zero. We have already seen, in the example with classical music in Chapter **??**, that a main effect, such as age, would not be set to zero if an interaction of age with another explanatory variable, there education, is required in the model. This is the idea of hierarchical models.

A more difficult idea is that the parameters for margins corresponding to explanatory variables in logistic and log linear models must not be set to zero. For example, for log linear models, the minimal model is Equation (2.16) if all three variables are responses or only one is explanatory, whereas it is Equation (2.15) if those indexed by i and j are explanatory and that by k the response. None of the parameters in the appropriate one of these equations can be set to zero. For a logistic model, such as Equation (2.6), only μ could generally not be set to zero.

3.4 Calibrating the likelihood

Ways of calibrating the likelihood function are one of the hottest areas of debate in modern statistics. Students should probably be told up front that different statisticians they meet may give them conflicting advice on this subject.

3.4.1 Degrees of freedom

Here, we come back to the idea of degrees of freedom, the students discovering that the number of parameters in a model for dependence in a two-way table corresponds to the number of arbitrary entries when the margins are fixed.

Goodness of fit is a difficult concept that should be clearly shown to involve two conflicting criteria. A model should be close to the observed data. But, at the same time, it should not be too complicated. The incommensurability of complexity and measures of closeness to the data should be emphasised, leading into the problem of calibrating the likelihood function.

3.4.2 Model selection criteria

In contrast to significance tests, to be presented next, students rightly love the AIC. This simple weighting of closeness to the data, the deviance, with complexity of the model, the degrees of freedom, is intuitively appealing and easily applicable. However, it is probably a good idea to mention that the weighting of two times the number of parameters is arbitrary, although standard. To obtain simpler models, the penalty should be increased; this must be decided before starting to analyse the data.

Thus, in my experience, once students have seen the simplicity of the AIC, they prefer it to any other method, no matter how hard I try to push the other approaches. No tables required!

3.4.3 Significance tests

Trying to explain the rationale of significance tests is probably the most difficult task of the course. The students already know that a large deviance indicates that the simpler model is considerably less acceptable in the light of the data, strictly in terms of how close it is to those data. Now, they learn that large deviances will be rare if the simpler model is ‘correct’. (Correct is in quotes because this is a fiction; no model is ever correct, but only a rough simplification of reality.) Because the probability of large deviances is usually difficult to calculate, an approximation is called for: the Chi-squared distribution. (It was a sociology student who asked me why, if the preceding argument was correct, small deviances are rare.)

3.4.4 Prior probability

The idea of prior probabilities is also easy to get across. However, in realistic problems, it is much more difficult to implement. How can prior probabilities for all of the parameters of even a simple logistic model be derived, especially to be coherent under different constraints?

3.5 Goodness of fit

3.5.1 Global fit

Up until now, goodness of fit has been a very relative characteristic of a model, only being in comparison to another model. The existence of a saturated model yields a more absolute criterion because it fits the data exactly. Slipping in the comparison of the smooth density function to an empirical histogram is an important lead up to the next chapter. Here, the relationship between model smoothing and simplification can be clearly brought out. The conclusions for the logistic model fitted to the malformed children data confirm that the smoothing of these data is useful: there is little or no evidence in these data of a beneficial affect of a little alcohol!

3.5.2 Residuals and diagnostics

For the decomposition of goodness of fit, the well-known Pearson residuals were chosen, instead of the more logical deviance residuals. The former allow the Pearson goodness of fit statistic to be introduced. The latter can also be presented to more sophisticated classes.

3.6 Sample size calculations

Sample size calculations should be an essential part of the planning of any study. Power calculations are far too complex for this level of course, and impossible to justify. In addition, they are only approximate for categorical data models whereas the approach used here is exact (although the formula given involves a slight approximation). For most classes, it may be sufficient to present the reasoning based on Figure 3.5 .

3.7 Solutions to the exercises

Question (1)

The Poisson distribution (Section 4.2.3) is given by

$$\Pr(y_i) = \frac{\mathrm{e}^{-\mu}\mu^{y_i}}{y_i!} \qquad y_i = 0, 1, 2, \ldots$$

The maximum likelihood estimate of the mean is $\hat{\mu} = \frac{1}{N}\sum y_i$. Suppose that this is calculated to be $\hat{\mu} = 10$ with $N = 20$ observations.

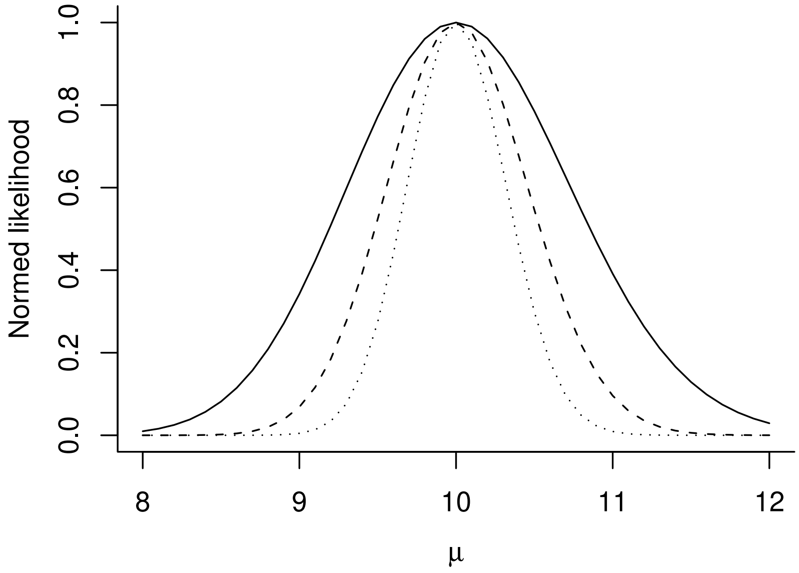

Figure 3.1: Normed likelihood functions for the mean of a Poisson distribution with $\hat{\mu} = 10$ and $N = 20$ (solid), 50 (dashed), and 100 (dotted).

(a) Plot the normed likelihood function.

(b) Repeat for the same estimate but $N = 50$ and $N = 100$.

Answer

(a) and (b) The normed likelihood functions are plotted in Figure **??**. We see how the graph becomes narrower as the sample size increases and we obtain more information. For example, the 10% likelihood intervals for μ are (8.55, 11.6) for $N = 20$, (9.05, 11.0) for $N = 50$, and (9.35, 10.7) for $N = 100$.

Question (2)

Calculate the AICs under independence for the logistic models which were fitted to the following data in the Exercises of Chapter **??**:

(a) the data on injuries in car accidents and wearing a seat belt of Table **??**;

(b) the data on myocardial infarction and contraceptive use of Table **??**;

(c) the data on opinions about the death penalty and gun registration of Table **??**.

In each case,

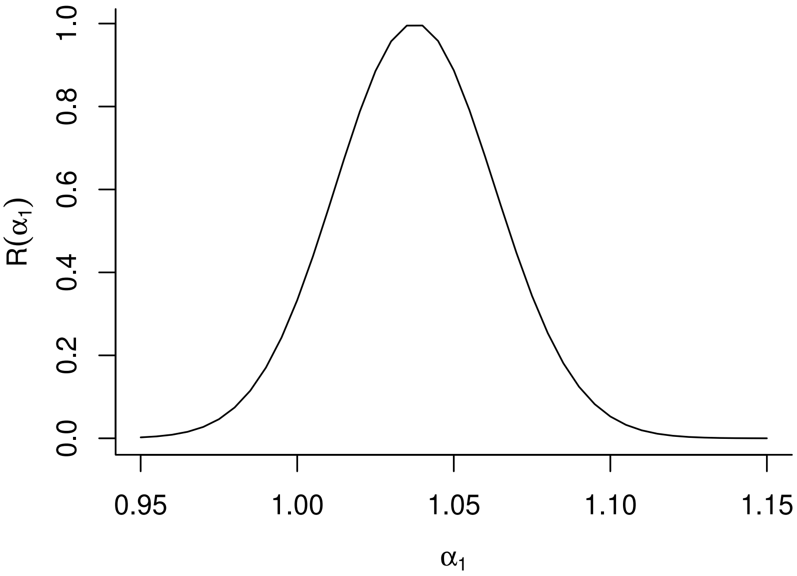

Figure 3.2: Normed likelihood functions for the dependence of a fatal accident on wearing a seat belt.

(a) plot the normed likelihood function for the dependence parameter and choose an interval of precision;

(b) discuss the conclusions which can be drawn;

(c) note whether they change anything which you concluded in Chapter **??**.

Can we conclude

(a) that making seat belts compulsory will reduce the fatal accident rate?

(b) that using contraceptives is a cause of myocardial infarction?

Answer

(a) For the car accident data, the AIC for independence is 2043.2 as compared to 4.0 for the saturated model. The normed likelihood function is plotted in Figure **??**. A 0.2 normed likelihood interval is rather narrow, about $(1.00, 1.08)$, with maximum likelihood estimate 1.038.

We can clearly conclude that the probability of a fatal accident is greatly increased when a seat belt is not worn. However, this does not mean that we can also conclude from these data that wearing a seat belt protects the person concerned. For example,

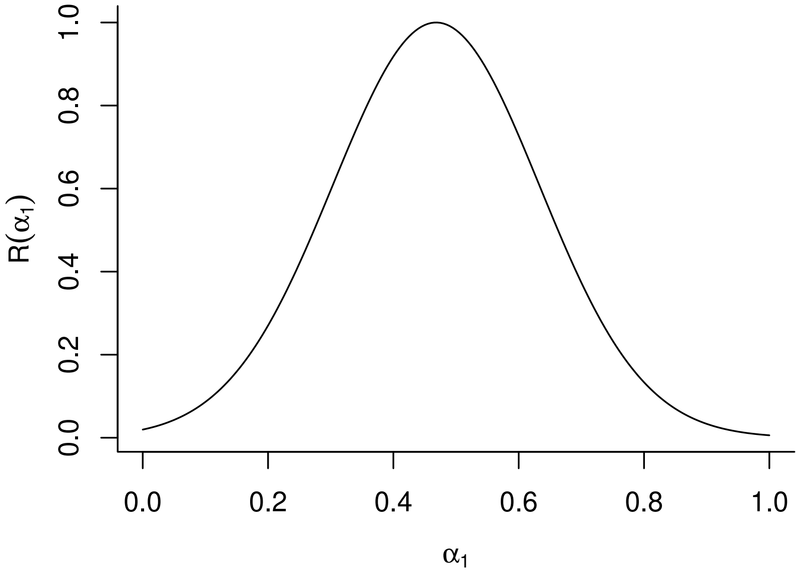

Figure 3.3: Normed likelihood functions for the dependence of a myocardial infarction on contraceptive use.

as mentioned in Chapter **??**, more careful drivers may wear a seat belt, but have less serious accidents simply because of their care.

(b) For the myocardial infarction data, the AIC for independence is 9.9 as compared to 4.0 for the saturated model. The normed likelihood function is plotted in Figure **??**. A 0.2 normed likelihood interval is here much wider, about $(0.17, 0.77)$, with maximum likelihood estimate 0.468. However, it does not cover 0.

We can conclude that the higher rate of myocardial infarction when contraceptives have been taken is unlikely actually to be the same as without them. Again, this does not mean that we can draw a causal conclusion. The women who decide to take contraceptives might have more inherent susceptibility to such attacks.

(c) For the death penalty and gun registration data, the AIC for independence is 7.3 as compared to 4.0 for the saturated model. The normed likelihood function is plotted in Figure **??**. A 0.2 normed likelihood interval is wider still, about $(-0.08, -0.01)$, with maximum likelihood estimate -0.087. It is close to covering 0.

Again, we reject independence, but somewhat less strongly than before. Apparently, there is a non-zero association between being in favour of the death penalty and against gun registration.

Notice that the AIC for the saturated logistic model, allowing for dependence, is 4 in all cases. For saturated models, the deviance is 0.

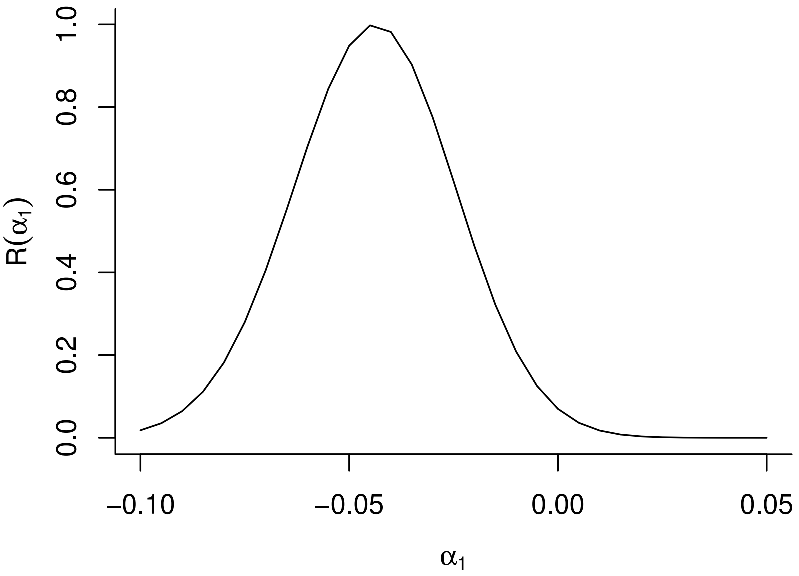

Figure 3.4: Normed likelihood functions for the association between the death penalty and gun registration.

Question (3)

Calculate the AICs under independence for the logistic models which were fitted to the following data in the Exercises of Chapter **??**:

(a) the data on soloists' choice of piano in Exercise (**??**.**??**);

(b) the British migration data of Table **??**.

In each case,

(a) discuss the conclusions that can be drawn;

(b) note whether they change anything which you concluded in Chapter **??**.

Answer

(a) For the soloists' choice of piano, the AIC is 8.7 as compared to 12.0 for the saturated model. Although the probabilities of using a Steinway in the various orchestras appeared to be different in Exercise (**??**.**??**), this indicates that there is no evidence of them actually being different

(b) For the migration data, the AIC is 19888.1 as compared to 32.0 for the saturated model. Here, with this extremely large value, there is clear evidence that the place of residence in 1971 is not independent of that in 1966. As we saw in the previous chapter, almost all of this dependence is due to the people who are in the same place at the two dates. Another way to tackle this data set would be to study the relationship between the origin and the destination only of the observed movers, ignoring the diagonal.

Question (4)

Calculate the AICs for independence of the response from the explanatory variables for the logistic or log linear models which were fitted to the following data in Chapter **??**:

(a) the tables of Exercise (**??**.**??**) concerning listening to the radio;

(b) the delinquency data of Exercise (**??**.**??**);

(c) the data on factors influencing knowledge of cancer in Table **??** and Exercise (**??**.**??**);

(d) the shopping data of Exercise (**??**.**??**).

In each case,

(a) select one or more parameters in which you are especially interested, plot their normed profile likelihood function(s), and choose appropriate intervals of precision;

(b) discuss the conclusions which can be drawn;

(c) note whether they change anything which you concluded in Chapter **??**.

Answer

(a) The AIC for independence of listening to religious programmes from age and education is 157.8 whereas that for discussion programmes is 72.2, as compared to 8.0 for both saturated models. In both cases, there is strong evidence that listening to these programmes depends on age and/or education. The AICs when listening depends only on age are respectively 136.0 and 76.3, whereas those with only education are 6.3 and 50.5. Thus, listening to religious programmes appears to depend only on education whereas listening to discussion programmes depends on both. In the latter case, the interaction is not necessary.

Lower educated people listen more to religious programmes ($\widehat{\alpha_2} = 0.693$). Higher educated and older people listen to discussion programmes ($\widehat{\alpha_1} = 0.313$ and $\widehat{\beta}_1 = 0.308$). The normed likelihood functions are plotted in Figure **??**. For religious programmes, a 0.2 normed likelihood interval for the education parameter is about $(0.59, 0.81)$ and for age $(-0.02, 0.18)$. The latter contains 0. For discussion programmes, they are respectively $(0.23, 0.39)$ and $(0.23, 0.40)$.

(b) For the delinquency data, the AIC for independence from SES and being a boy scout is 34.8 as compared to 12.0 for the saturated model. However, we can also fit for the marginal table with only SES, where we find the same deviance as the saturated model, but an AIC of 6.0. The normed likelihood function for the boy scout parameter is plotted in Figure **??**. A 0.2 normed likelihood interval is about $(-0.22, 0.22)$ clearly covering 0. This indicates that there is no evidence that delinquency depends on being a boy scout, but only on SES.

(c) The AIC for independence of knowledge of cancer from the three sources in Table **??** is 127.8, whereas, when lectures are replaced by reading, as in Exercise (**??**.**??**), it is 203.0, both as compared to 16.0 for the saturated models. This reflects the fact that reading is much more strongly related to such knowledge than are lectures. In this case, the problem illustrated by Figure 3.2 of the text does not occur. The estimate for the effect of reading is further from zero than that for lectures, but the likelihood curve must not be much wider (if at all).

For the complete table with all four explanatory variables, the AIC for independence is 218.3 as compared to 32.0 for the saturated model. After simplification, with three interactions left, the AIC is 20.3. Because of the interactions, there is no particular parameter of special interest for which a likelihood function might be plotted.

(d) For the shopping data, the AIC for independence is 16.1 as compared to 16.0 for the saturated model. Not much to choose between! Likelihood functions might be plotted for any of the dependence parameters. All will easily include 0.

Question (5)

As in the previous question, calculate the AICs for independence of the response from the explanatory variables for the polytomous logistic model for the following data from Chapter **??**:

(a) the political ideology data of Exercise (**??**.**??**);

(b) the media rating data of Exercise (**??**.**??**).

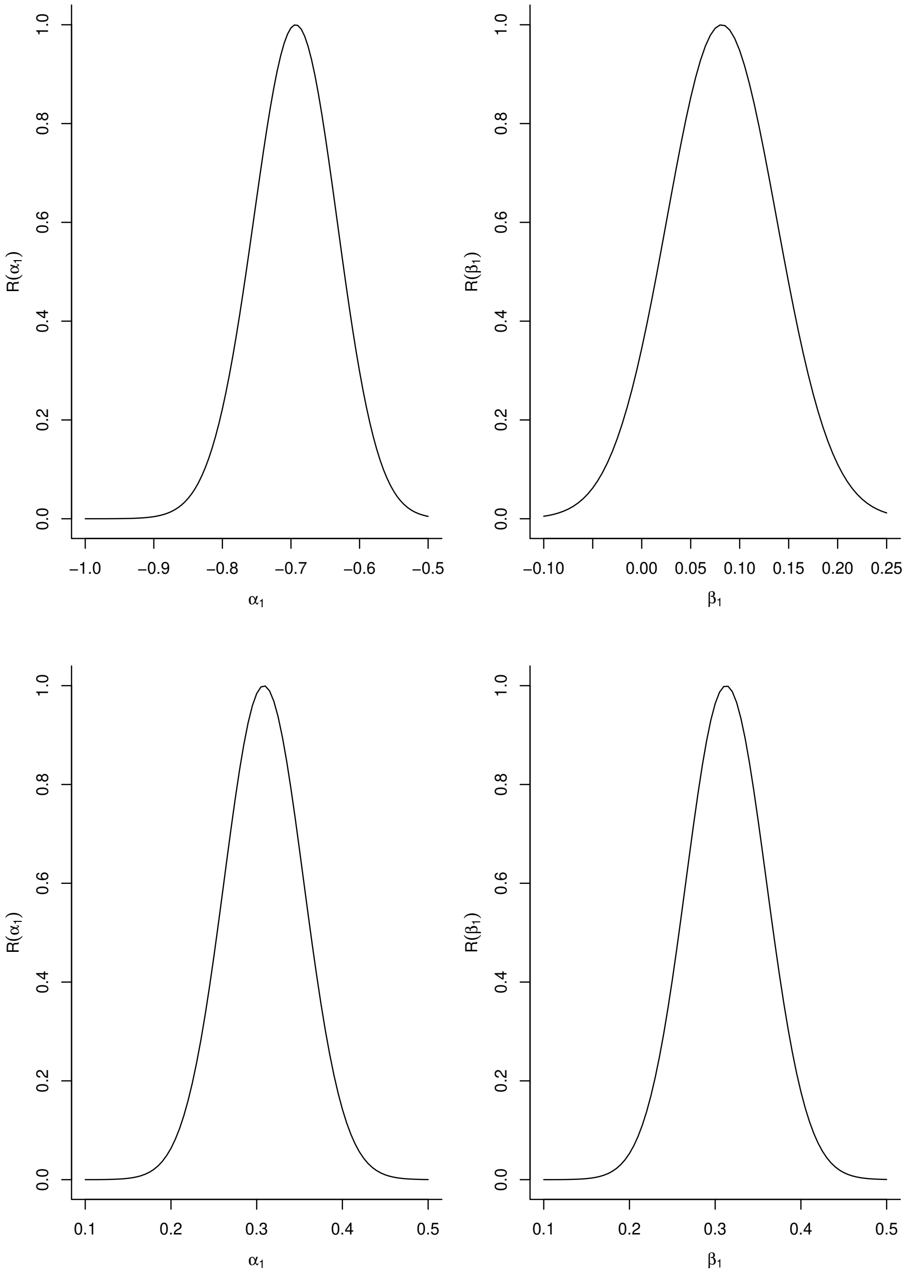

Figure 3.5: Normed likelihood functions for the dependence of listening to religious (top) and discussion (bottom) programmes on education (α_1) and age (β_1).

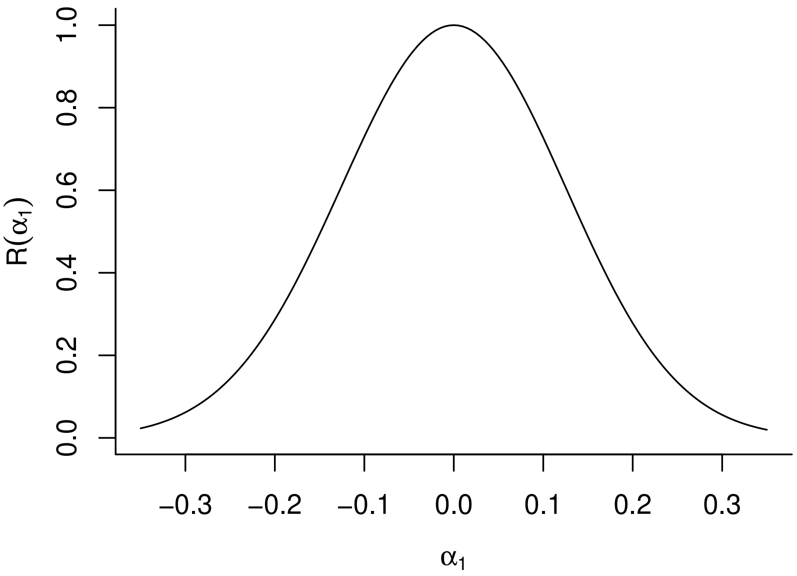

Figure 3.6: Normed likelihood functions for the dependence of a delinquency on being a boy scout.

In each case,

(a) redo the AIC calculations for the reconstructed table for the two ordinal variable models;

(b) compare the results and discuss the meaning of any differences.

Answer

(a) For the political ideology data, the AIC for independence of party from ideology in the original table, not allowing for an ordinal response variable, is 115.7 as compared to 18.0 for the saturated model. With the table for the continuation ratio model, the AIC is 238.1, whereas with that for the proportional odds model, it is 515.5, both as compared to 12.0 for the saturated models. The models without interaction have respectively AICs of 9.2 and 11.0. Note that only AICs for the same table are comparable!

The conclusion is clear: the model with independence between ideology and party is not supported.

(b) For the media rating data, the AIC for independence of the rating from both colour and year in the original table is 77.7 as compared to 24.0 for the saturated model. The deviances for the two reconstructed tables are 475.9 and 1201.3, both as compared to 16.0 for the saturated models. The models without interaction between the subtables and the other two variables have respectively AICs of 13.3 and 10.3. When the interaction between colour and year is removed, the AIC increases in both cases.

We have evidence that rating depends both on colour and on year.

Question (6)

Suppose that the shopping data of Exercise (**??**.**??**) were collected by a firm considering the construction of a new shopping centre in the same region.

(a) Specify an appropriate null hypothesis for making such a decision. (Try to do this without using what you already know about these data!)

(b) Calculate the corresponding test of significance.

(c) Describe what subsequent action you would advise should be taken.

Answer

(a) Three possibilities are that choice of shopping centre does not depend on availability of a car, that it does not depend on age, and that it does not depend on income. In the first case, if people with a car come more often, a large parking lot should be planned. In the second case, if say young people are to be attracted, suitable shops should be made available. In the third, if high income people are more frequent, high class shops should be located in the centre.

(b) The corresponding deviances are respectively 8.34, 3.99, and 3.91, all with 4 degrees of freedom. The respective P-values are 0.08, 0.41, and 0.42 from a Chi-squared test. Hence, none of the null hypotheses are rejected at the 5% level.

(c) People in the area do not seem to have specific habits as to shopping so that a generalised shopping centre might be more suitable than a specialised one.

Question (7)

Suppose that the study on myocardial infarction and contraceptive use reported Table **??** was conducted to decide whether or not to withdraw contraceptives from the market.

(a) Based on the normed likelihood function that you plotted in Exercise (**??**.**??**) above, calculate an appropriate confidence interval for the dependence of myocardial infarction on contraceptive use.

(b) What recommendations would you make to the policy deciders?

(c) Suppose now that you believe that only two values of the dependence parameter could reasonably be true, one of them being that for independence and the other being a log odds ratio of one.

  i Assign your prior probabilities to these two possibilities. (Try to do this without using what you already know about these data!)

  ii Obtain the updated posterior probabilities.

  iii Has your opinion on the subject now changed and, if so, in what way?

Answer

(a) A 95% interval will have a deviance of 3.84 or a normed likelihood of $\exp(-3.84/2) = 0.14$. From the graph, this interval is about $(0.96, 1.09)$ with maximum likelihood estimate 1.038.

(b) The model without dependence is clearly rejected. The dependence of myocardial infarction on contraceptive use should clearly be investigated further.

(c) If I believe that either is equally possible, my prior probabilities are both 0.5. The probabilities of the observed data under the two hypotheses are 0.000492 under independence and 0.002218 with log odds ratio of unity. The posterior probabilities are then 0.18 for independence and 0.82 for the unit log odds ratio.

(d) Thus, my indifference changes to a marked preference for the latter.

Question (8)

(a) What size of sample would be required to detect that a log odds ratio was 1.0 as opposed to zero, with a deviance of at least 5? Assume, as in the example above, that you can choose a sample with equal numbers in each category of the explanatory variable.

(b) Plot the required sample size for several values of the mean probability of response.

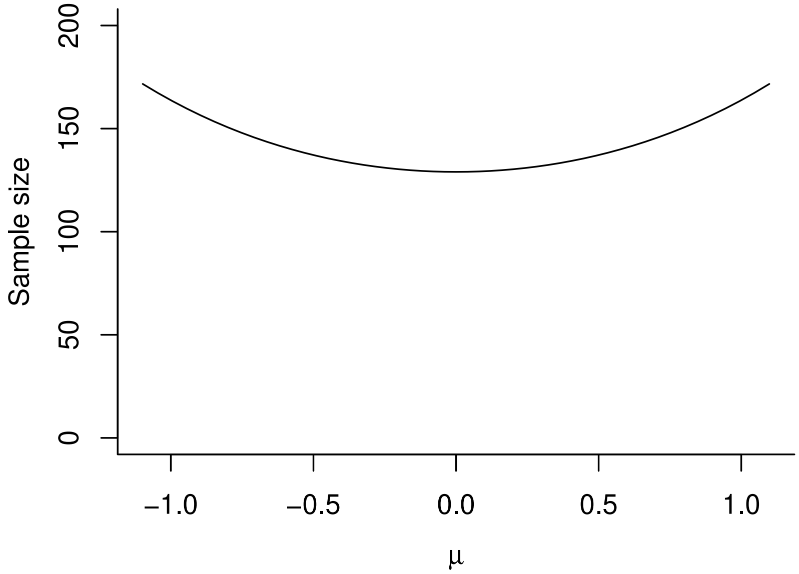

Figure 3.7: Sample size calculations to detect a log odds ratio of one as opposed to zero, for different values of the mean parameter.

Answer

(a) The value of α_1 for a log odds ratio of 1.0 is 0.25. If the mean parameter, μ is assumed to be zero, the required sample size is $N = 129$. We require a larger sample than for the example in the text because the difference in log odds ratio is smaller.

(b) For various other values of the mean parameter, the required sample size is plotted in Figure **??**. We see how the required sample grows as the mean moves away from zero.

Chapter 4

Probability distributions

Here we come to the second chapter (after Chapter **??**) that is meant to convince students that statistics can really have some practical use. When the course is finished, mine generally judge this to be the most interesting chapter. The basic idea is simple and can become repetitive: choose some distribution that might appropriately describe the data generating mechanism, calculate the estimated probabilities from its function, compare them to the relative frequencies, and draw the relevant conclusions.

4.1 Constructing probability distributions

4.1.1 Multinomial distribution

This is just a review of material from Chapter **??**.

4.1.2 Density functions

Density functions were briefly discussed at previous points in the course. Here, we really get down to business. One of the main jobs of statistics is smoothing random variation to find interesting patterns in it. This is what density functions do for response variables.

The basic idea in constructing discrete probability functions is simple: normalise any set of positive values by dividing by their sum so that the members of the series fulfil the two criteria of a probability. Most classes will have to accept on faith that infinite series can be summed.

Here, the width of histogram bars in Chapter **??**, Δ_i, becomes the unit of measurement. This can be confusing because it is not necessarily the units in which the measurements are expressed, but their precision. Thus, a measurement may be in metres, but given to the nearest five metres or to one tenth of a metre. In the first case, $\Delta_i = 5$ and, in the second, $\Delta_i = 0.1$.

For the calculation of parameter estimates, generally the centres of intervals can be used. Problems can occur with open intervals for the extreme values of a variable.

No simple solution is possible at this level (i.e. without using censoring), so that some reasonable value must be chosen.

Students easily acquire the idea that there is a correct distribution for every data set and that their job is to find it. This illusion should be dispelled as quickly as possible. Distributions, and all models, are simplified scientific representations of reality. The ones chosen for inspection for a given data set will depend on the questions being asked, and should, ideally, be chosen for theoretical reasons related to the suspected data generating mechanism, although such theories are not always available. In addition, the fact that no distribution is found to fit to the data can be very informative: those tried as reasonable possibilities and rejected indicate that the corresponding mechanisms are likely not to be working.

I have divided distributions into a few main classes according to their primary domains of application. Students should be shown how this provides a first criterion for reduction of the number of possible distributions under consideration for modelling a given data set. Obviously, the classification is not inviolable. Days can be counted, although they are divisible, so that, in certain contexts, one can reasonably argue that count distributions can be applied to a response measured in days.

4.2 Distributions for ordinal variables

4.2.1 Uniform distribution

The uniform distribution provides a gentle but useful introduction to the fundamental ideas of the chapter. The basic assumption, constant probability, is simple. Make sure that the students understand how to calculate the AIC for the saturated multinomial model: $2 \times (I-1)$. This will provide the basis of comparison throughout the chapter.

4.2.2 Zeta distribution

The zeta distribution is interesting in itself for its applications in linguistics. The example is rather artificial, but provides the opportunity to introduce the idea of a truncated distribution. In this way, the students should understand more clearly how probability distributions are constructed.

4.3 Distributions for counts

4.3.1 Poisson distribution

The Poisson distribution is traditionally introduced as representing rare events. This is not very useful for modelling. On the other hand, randomness is an important characteristic beneath all statistics. This distribution allows us to detect it, or, more often, to conclude that it is not likely present. The underlying concept is the Poisson process which should be discussed. Counts are generally aggregations of events over time. How much and what type of information is lost by aggregation?

The fact that the variability of a distribution can change with the mean is an important relationship. Unfortunately, it is often ignored in classical statistics because the normal distribution does not have this relationship. The Poisson distribution provides a good opportunity to emphasise its importance. A graph of several Poisson distributions with different means readily shows that those with means closer to zero must pile up, because of the restriction that values cannot be negative, hence have less variability.

The coefficient of dispersion provides a simple indicator of lack of randomness. However, its limitations, especially when close to one, should be presented. The fact that the variance is similar in size to the mean does not imply a Poisson distribution. Other distributions might have this characteristic (e.g. normal). One really must look at the whole shape of the distribution in order to learn the most about the phenomenon under study.

The example of ages of children in Bombay is meant to provide a realistic (and true) case of model building, starting from the first lesson of the course: construction of an appropriate variable. The result is a rather complex model, although the mathematics are minimal. The shape of the distribution and the mean parameter, as each varies among the classes, provide different aspects of a possible answer to what is going on. The need for data on the individual trajectories of children (a point process, possibly Poisson) in order to check such hypotheses should be discussed in detail with the students. Why is such information necessary to verify what we expect is happening?

The relationship between the Poisson and multinomial distributions is only really useful if the students will be continuing on to a more advanced categorical data course where it will be used. Otherwise, it can be skipped.

4.3.2 Geometric distribution

The geometric distribution provides the first opportunity to introduce the Markov property. Like the randomness of the Poisson distribution, this ‘lack of memory’ model provides a valuable ‘null’ model that will most often be rejected.

The first example introduces a first approach to duration data.

The second example for this distribution provides a new occasion to show that the obvious variables are not necessarily the best. If the students have not read ahead, they should be asked to try to figure out how the variable recording the number of occupants of vehicles could be adapted to make a variable appropriate for the geometric distribution.

4.3.3 Binomial distribution

The binomial distribution presents a good occasion to examine what modelling assumptions can really imply. Is the probability of a boy the same in all families? Does it remain the same in successive births in the same family? This may be usefully compared to applying the same model to students correctly answering successive questions on a test: questions are more variable (in difficulty) than children, at least as far as the probabilities involved are concerned.

Before introducing the formula for the maximum likelihood estimate of ν_1, get the students to figure out intuitively what the estimate of the probability of a boy should be for Table 4.8. They almost certainly will come up with the correct formula, although not necessarily in mathematical terms.

This example provides an extremely clear case for the usefulness of residuals. This first introduction of overdispersion is meant to lead to the search for reasons for its presence. Thus, possible explanations for the excess of families with the majority of children of one sex should be discussed with the students. Point out that this cannot even be detected without the underlying assumptions of the binomial model. Simple models can provide useful results, even when they are wrong.

4.3.4 Negative binomial distribution

This common distribution for overdispersion with the Poisson distribution, the negative binomial, is also useful in its own right. It is another distribution for which maximum likelihood estimates cannot easily be obtained so that an approximation is used if the course is not based on computers. Note that the ratio of gamma functions can be simplified by calculation so that a table of values is not necessary.

Although there is an extreme number of people not buying any items in the example, the negative binomial distribution manages to fit rather well.

4.3.5 Beta-binomial distribution

For the example with the binomial distribution, we discovered overdispersion. Unfortunately, the common distribution to handle this, the beta-binomial, is relatively complex and the lecturer may want to skip it.

4.4 Distributions for measurement errors

4.4.1 Normal distribution

The normal distribution is useful for having the students practise calculating probabilities as the area under the density curve. Through the use of tables, they can see that this is possible even if the distribution is not based on a variable with discrete categories. Perhaps you will have more luck than I have in finding a realistic application of the normal distribution.

4.4.2 Logistic distribution

The logistic distribution is introduced for two reasons. It provides a first model with thicker tails than the normal and it provides one underlying justification for the logistic models of Chapter **??**. However, it, and the following three distributions, are not fundamental, and can be omitted except in fairly advanced classes.

4.4.3 Laplace distribution

Although historically important, this distribution is remarkably underused. It has a simple form, easily obtained estimates, and very often provides a better fit to data than the normal distribution.

4.4.4 Cauchy distribution

I included the Cauchy distribution in the 1970s version of this course, but omitted it from the first edition of this book. However, it is worth presenting to students, among other reasons, because it has no mean or variance!

4.4.5 Student t distribution

Because it has three parameters, this is one of the most complex distributions of the chapter. But this also means that it is a flexible distribution for difficult data sets. For those emphasising testing, it will useful be in Chapter **??**.

4.5 Distributions for durations

I have called this group duration distributions because of their most important application. However, it should be made clear to students that this does not mean that they can only be applied to durations. They will be of interest for any response that can only take positive values, especially so if many of the observed values are close to zero.

Duration distributions are one of the areas where the greatest advances have been made in modern statistics. However, they have much wider application; in most cases when continuous responses are recorded, one of these asymmetric distributions will be more appropriate than the normal distribution traditionally used.

4.5.1 Intensity and survivor functions

The basic tools of survival analysis are too important not to be presented in an introductory statistics course.

The idea of the rate or intensity of events is essential to get across. If the intensity, i.e. the mean number per unit time, is greater, then the average time between events will be shorter. Thus, the two are reciprocally related.

4.5.2 Exponential distribution

Randomness of events is connected to the memorylessness Markov property as to when the next will occur. I give the example of waiting for the bus to go to class on a cold snowy day: no matter how long you wait, the probability of a bus arriving does not increase!

4.5.3 Weibull distribution

The Weibull distribution is a flexible and widely used model with nice properties (the cumulative distribution and intensity functions can be written down in closed form). Thus, it is important for analysing censored survival data, being one of the proportional hazards models.

4.5.4 Gamma distribution

The gamma distribution is one of the most important duration distributions (with the exponential, Weibull, log normal, and inverse Gauss). It is especially interesting when α is close to being an integer so that it can be interpreted in terms of the sum of several periods. Note that here, in contrast to the negative binomial distribution, if a computer is not used, a table is necessary to obtain the value of the gamma function.

4.5.5 Inverse Gauss distribution

The inverse Gauss distribution is one of the most under-used in statistics. Its development, as based on Brownian motion, can have many applications, limited only by one's imagination. This provides a further occasion to emphasise how most responses evolve over time. Here, the whole theoretical system, as depicted in Figure 4.31, is strictly theoretical, with only the final event, the border crossing being observed. Students generally find this a realistic and interesting model and are ready to look for applications.

4.6 Transforming the response

Transformations of variables had an important role before computer technology allowed a wide choice of distributions. They played a number of roles, but the essential one in realistic modelling is to change the shape of the response distribution, creating a new distribution.

4.6.1 Log transformation

The log normal distribution may be the most widely used model for continuous positive data. Most of the applications of the normal distribution, such as in economics, are in fact log normal because a log transformation has been applied to the response variable. The students may be induced to discuss when a large number of factors might multiply or add together to give a response.

The Pareto distribution has historical interest in economics and sociology, but does not have too many practical applications.

4.6.2 Exponential transformation

As its name indicates, the extreme value distribution is central to the study of extremes. This should be developed further than in the text if your students will have such applications.

4.6.3 Power transformations

The Box–Cox model is only mentioned in passing for completeness. Generally, it will only be discussed in fields such as economics where it is still used. In fact, it is not even a real probability model because the function produced does not integrate to unity!

Table 4.1: Numbers of infants born with an illness each month of a year.

Month	Number
January	8
February	19
March	11
April	12
May	16
June	8
July	7
August	5
September	8
October	3
November	8
December	8
Total	113

4.7 Special families

4.7.1 Location-scale family

This family shows the unity of the distributions presented in Section **??** on measurement error.

4.7.2 Exponential family

The theory of the exponential family was one of the bases of modern parametric statistics because of its role in inference. However, this section will be too advanced for most students who do not have a fair bit of mathematics behind them. It is not essential to modern statistical modelling.

4.8 Solutions to the exercises

Question (1)

In the text, I fitted a uniform distribution to the numbers of ill children born each month, given in Table **??**. My model did not take into account the fact that months have different numbers of days.

(a) Construct a new model, based on the uniform distribution, using this information.

(b) Does it fit better than the previous model?

Answer

(a) The new model will have theoretical probabilities, not of 1/12 for each month, but of the number of days in each month divided by 365: (31, 28, 31, 30, 31, 30, 31, 31, 30, 31, 30, 31)/365. (We do not know what year it was so assume that it was not a leap year.) We found a deviance of 22.47 for the first model. The second one has 24.42. The AIC is the same as this because again there are no estimated parameters.

(b) Thus, this more accurate model fits more poorly. However, we must prefer it to that using 1/12. If we look at the data, we see that the reason it fits more poorly arises primarily from February, the shortest month which has the most ill children. Thus, this better model provides more evidence that the probability of ill children is not constant over the year.

Question (2)

The following table gives the frequency of occurrence of surnames from a study area of Reading, Workingham, and Henley-on-Thames, England (Fox and Lasker, 1983). The names for the complete study were those of all 2397 couples whose marriages were registered in the study area during a twelve month period in 1972–1973. Those given in the table are for one of eight districts of that area. Although the sample may have been reasonably representative of surnames in that geographical area, it is clearly not random with respect to age or other characteristics. As well, some of the people may only have been in the area for the purpose of registering their marriage.

Number of occurrences	Number of surnames
1	329
2	43
3	11
4	1
5	0
6	1
7	0
8	0
9	1

(a) Choose an appropriate probability distribution and fit it.

(b) Calculate the AIC and check the residuals.

(c) Discuss how well the model fits.

(d) What general conclusions could be drawn, given the way in which the sample was selected?

Answer

(a) An appropriate distribution might be the zeta. However, when we fit it as in the example in the text, with $\rho = 1$, the most common name is extremely underestimated. This indicates that we require a larger value of ρ. Without having an appropriate

Table 4.2: Frequency of surnames, with fitted zeta distribution and standardised residuals. (Fox and Lasker, 1983)

Occurrences	Names	Multinomial	Zeta	Residual
1	329	0.852	0.836	0.356
2	43	0.111	0.104	0.421
3	11	0.028	0.031	−0.274
4	1	0.003	0.013	−1.800
5	0	0.000	0.007	−1.606
6	1	0.003	0.004	−0.404
7	0	0.000	0.002	−0.970
8	0	0.000	0.002	−0.794
9	1	0.003	0.001	0.838

method to estimate it, we can try successive integers. We soon find that $\rho = 3$ provides a good fit. (The maximum likelihood estimate must be close to this value; in fact, it is 3.29.) The results are shown in Table **??**.

(b) The deviance is 14.23 and the AIC 16.23, as compared to 16 for the multinomial. The residuals show a systematic pattern, with underestimation of the first two categories. This could be improved by taking a slightly larger value of ρ (i.e. the maximum likelihood estimate, which will have a smaller AIC).

(c) The zeta model is plotted in Figure **??**. It can be seen to fit well, as indicated by the AIC and residuals. The relatively larger value of ρ required indicates that the distribution of surnames drops very quickly, i.e. that there is an extreme number of surnames with one couple and few with more.

(d) The people in this sample will be primarily in the age group twenty to thirty-five and the model may be useful to represent their distribution of surnames. However, factors such as differential birth rates over time, migration, and so on could greatly modify the distribution in other age groups and even in this one. For example, if the region has a number of single men who are migrant workers these will not be represented at all.

Question (3)

The table in Exercise (**??**) gave the frequency of burglaries in Detroit.

(a) Choose an appropriate probability distribution and explain your choice.

(b) Fit the distribution.

(c) Calculate the AIC and check the residuals.

(d) Discuss how well the model fits.

Answer

(a) We are interested to see if burglars strike randomly so we choose the Poisson distribution. The maximum likelihood estimate of the mean number of burglaries per

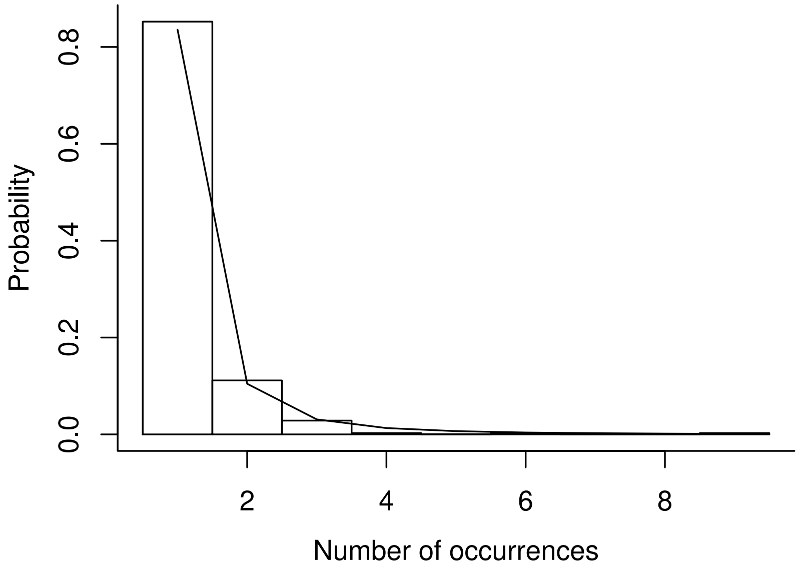

Figure 4.1: Histogram for surnames, with fitted zeta distribution. (Fox and Lasker, 1983)

Table 4.3: Frequency of burglaries in Detroit, with fitted Poisson and negative binomial distributions and standardised residuals. (Nelson, 1980)

		Multinomial	Poisson		Negative binomial	
Burglaries	Houses		Fitted	Residual	Fitted	Residual
0	8385	0.8747	0.8572	1.852	0.8741	0.064
1	976	0.1018	0.1321	−8.153	0.1033	−0.457
2	183	0.0191	0.0102	8.653	0.0181	0.694
3	35	0.0037	0.0005	13.399	0.0035	0.199
4	5	0.0005	0.0000	10.943	0.0007	−0.727
5	2	0.0002	0.0000	25.859	0.0002	0.456

household is $\hat{\mu} = 0.154$. The results for the fit of this distribution are given in Table **??** and plotted in Figure **??**.

(b) The deviance is 253.32 and the AIC 255.32, as compared to 10 for the multinomial distribution. The residuals were given in Table **??**.

(c) The Poisson distribution fits very badly. The estimated variance is $s^2 = 0.20$, considerably larger than the mean (coefficient of dispersion, 1.3), indicating some overdispersion. When we inspect the residuals, we see that households with one burglary are under-represented. This explains the overdispersion. Thus, there appear to be two groups of houses, those without burglaries, and those with multiple entries. If we fit the negative binomial distribution, we obtain a major improvement, with an AIC of 6.29. The parameter estimates are $\hat{\nu}_1 \doteq 0.77$ and $\hat{\gamma} \doteq 0.51$. The results are shown in Table **??**. There is now only a slight indication of too few observed one-time burglaries.

The negative binomial distribution can be derived from a Poisson distribution where the mean varies in the population, following a gamma distribution. (There are many other ways to describe overdispersed counts.) By allowing for the variability in the mean, this model can account for the two main groups, no burglaries and multiple burglaries (which have small and large means). Thus, much of the heterogeneity in the population with respect to burglary can be so described.

Question (4)

Let us reconsider the accident data in the two tables of Exercise (**??**.**??**)

(a) Choose appropriate probability distributions for each, explaining why.

(b) Fit them.

(c) Calculate the AICs and check the residuals.

(d) Discuss how well the models fit to each table.

Answer

(a) A simple hypothesis is that accidents happen at random, following a Poisson distribution. This ignores the variability in a population whereby some people are much

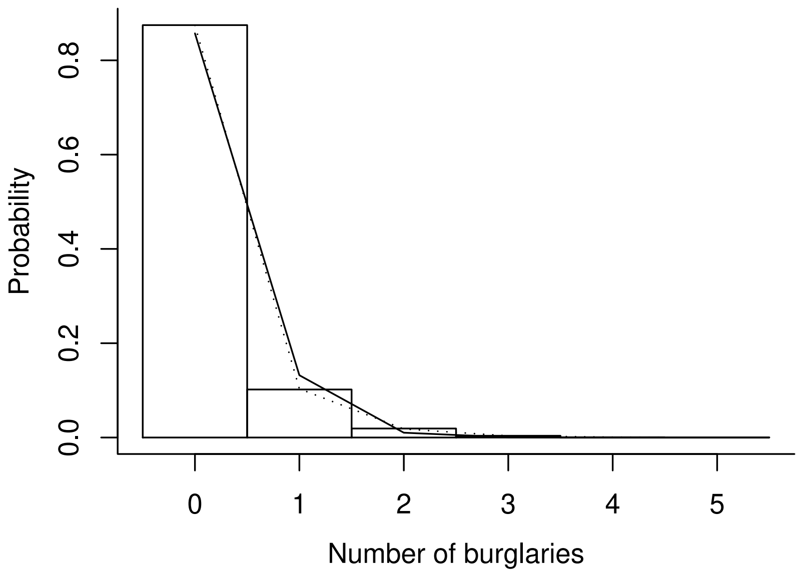

Figure 4.2: Histogram and fitted Poisson (solid) and negative binomial (dotted) distributions for burglaries in Detroit. (Nelson, 1980)

Table 4.4: Frequency of accidents, with fitted Poisson and negative binomial distributions and standardised residuals. (Skellam, 1948)

		Multinomial	Poisson		Negative binomial	
Accidents	People		Fitted	Residual	Fitted	Residual
0	447	0.691	0.6280	2.019	0.684	0.201
1	132	0.204	0.2922	−4.148	0.214	−0.570
2	42	0.065	0.0680	−0.297	0.069	−0.357
3	21	0.032	0.0105	5.431	0.022	1.773
4	3	0.005	0.0012	2.478	0.007	−0.754
5	2	0.003	0.0001	7.091	0.002	0.411

Table 4.5: Frequency of car accidents in Belgium, with fitted Poisson and negative binomial distributions and standardised residuals. (Gelfand and Dalal, 1990)

		Multinomial	Poisson		Negative binomial	
Accidents	Cars		Fitted	Residual	Fitted	Residual
0	7840	0.8287	0.8071	2.33890	0.8320	−0.353
1	1317	0.1392	0.1730	−7.90290	0.1323	1.841
2	239	0.0253	0.0185	4.80054	0.0276	−1.366
3	42	0.0044	0.0013	8.32299	0.0062	−2.185
4	14	0.0015	0.0000	16.26287	0.0014	0.079
5	4	0.0004	0.0000	23.40256	0.0003	0.407
6	4	0.0004	0.0000	124.68080	0.0001	3.616
7	1	0.0001	0.0000	178.16450	0.0000	1.841

more careful than others. The average number of accidents is estimated to be $\hat{\mu} = 0.47$ in the first table and $\hat{\mu} = 0.21$ in the second. The results for the models are displayed in Tables **??** and **??**.

(b) The deviances (AICs) are, respectively, 55.10 (57.10) and 302.48 (304.48). The latter compare with 10 and 14 for the multinomial AIC, indicating very poor fit. The residuals are large, with underestimation of small and very large numbers of accidents, confirming this impression. The estimated variances are slightly larger than the means, respectively $\bar{y}_\bullet = 0.47$, $s^2 = 0.69$ and $\bar{y}_\bullet = 0.21$, $s^2 = 0.29$ (coefficients of dispersion, 1.47 and 1.38), indicating some overdispersion.

(c) The models fit poorly so we try the negative binomial distribution which will allow for heterogeneity of accident proneness in the population. The parameter estimates are, respectively, $\hat{\nu}_1 \doteq 0.67$, $\hat{\gamma} \doteq 0.96$ and $\hat{\nu}_1 \doteq 0.74$, $\hat{\gamma} \doteq 0.62$. The results are shown in Tables **??** and **??**. The deviances and AICs are, respectively, 5.48 (9.48) and 19.22 (23.22) and the residuals are much smaller. The somewhat poorer fit for the second table results, at least in part, from the large number of observations which would require a more complex model to be described in a more adequate manner. The histograms, with both Poisson and negative binomial models, are plotted in Figures **??** and **??**. We see how the negative binomial distributions fit the data more closely.

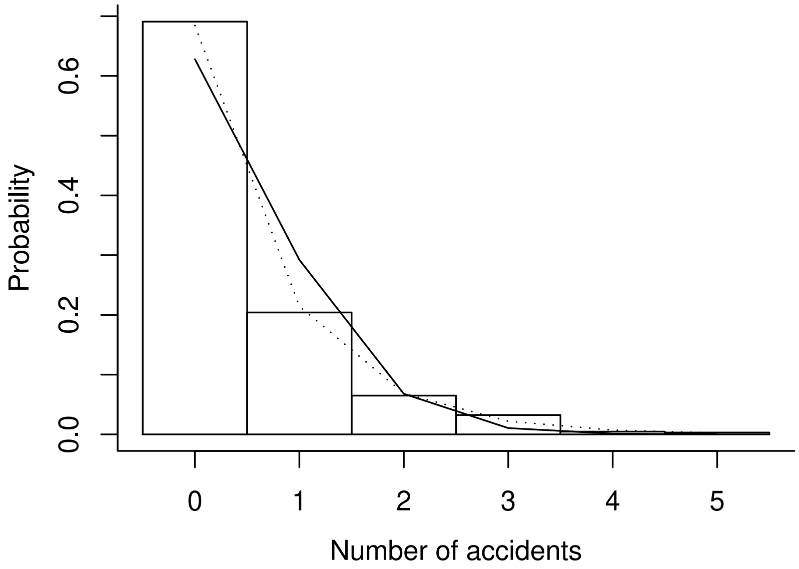

Figure 4.3: Histogram and fitted Poisson (solid) and negative binomial (dotted) distributions for the accident data. (Skellam, 1948)

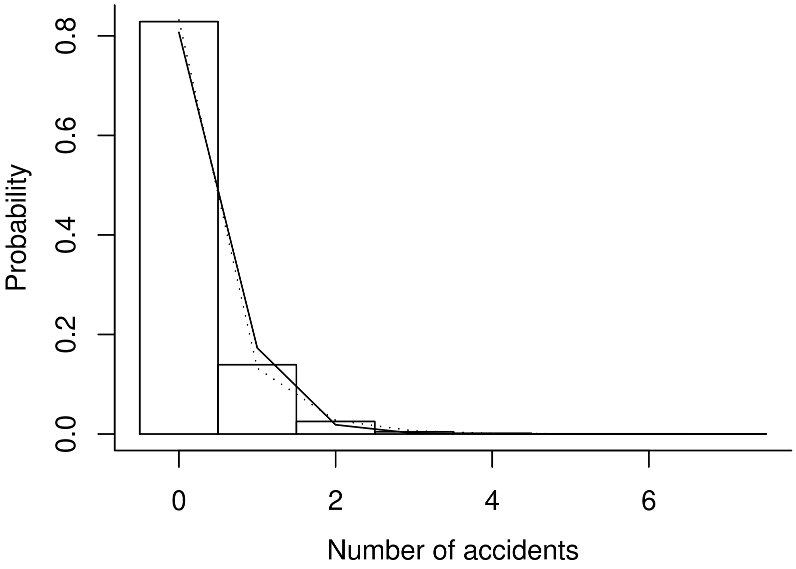

Figure 4.4: Histogram and fitted Poisson (solid) and negative binomial (dotted) distributions for the Belgian car accident data. (Gelfand and Dalal, 1990)

Table 4.6: Counts of accidents to men working in a soap factory over a five month period. (Irwin, 1975, from Newbold)

Number of accidents	Number of men	Number of accidents	Number of men
0	239	7	1
1	98	8	0
2	57	9	4
3	33	10	1
4	9	11	0
5	2	12	0
6	2	13	1

Question (5)

Table **??** gave the frequency of accidents to men working in a soap factory over a five month period.

(a) Choose an appropriate probability distribution, giving your reasons.

(b) Fit the distribution.

(c) Calculate the AIC and check the residuals.

(d) Discuss how well the model fits.

Answer

(a) Again, we wish to check if accidents are happening randomly, so that we use the Poisson distribution. The estimated mean number of accidents is $\hat{\mu} = 0.97$. The results are given in Table **??**.

(b) The deviance and AIC are 196.98 and 198.98, as compared to an AIC of 26 for the multinomial model. Some of the residuals are very large. The estimated mean number of accidents is $\bar{y}_\bullet = 0.97$ as compared to the estimated variance of $s^2 = 2.48$ (coefficient of dispersion, 2.56), clearly indicating overdispersion.

(c) Because the model fits poorly, we try the negative binomial. The parameter estimates are $\hat{\nu}_1 \doteq 0.39$ and $\hat{\gamma} \doteq 0.63$ and the results of model fitting are shown in Table **??**. The deviance is reduced to 29.00 with an AIC of 33.00, but this is still somewhat larger than that for the multinomial. The histogram, with both Poisson and negative binomial models, is plotted in Figure **??**. The sample size is not too large so that that is not the problem. Apparently, heterogeneity of proneness to accidents (at least with this model for overdispersion) is not sufficient to explain the overdispersion in these data. This might be due to important missing explanatory variables such as the type of job each worker is performing. For example, one could easily imagine a classification of the tasks in the factory as ‘administrative’ or ‘manual’. Different possibilities would arise then:

Table 4.7: Frequency of accidents in a soap factory, with fitted Poisson and negative binomial distributions and standardised residuals. (Irwin, 1975)

		Multinomial	Poisson		Negative binomial	
Accidents	Workers		Fitted	Residual	Fitted	Residual
0	239	0.535	0.3779	5.392	0.5553	−0.586
1	98	0.219	0.3677	−5.178	0.2122	0.322
2	57	0.128	0.1789	−2.570	0.1050	1.469
3	33	0.074	0.0580	1.385	0.0559	1.605
4	9	0.020	0.0141	1.070	0.0308	−1.284
5	2	0.004	0.0027	0.070	0.0173	−2.063
6	2	0.004	0.0004	4.034	0.0099	−1.147
7	1	0.002	0.0001	5.842	0.0057	−0.965
8	0	0.000	0.0000	−0.058	0.0033	−1.212
9	4	0.009	0.0000	209.531	0.0019	3.401
10	1	0.002	0.0000	16.793	0.0011	0.707
11	0	0.000	0.0000	−0.002	0.0007	−0.542
12	0	0.000	0.0000	−0.001	0.0004	−0.416
13	1	0.002	0.0000	724.641	0.0002	2.813

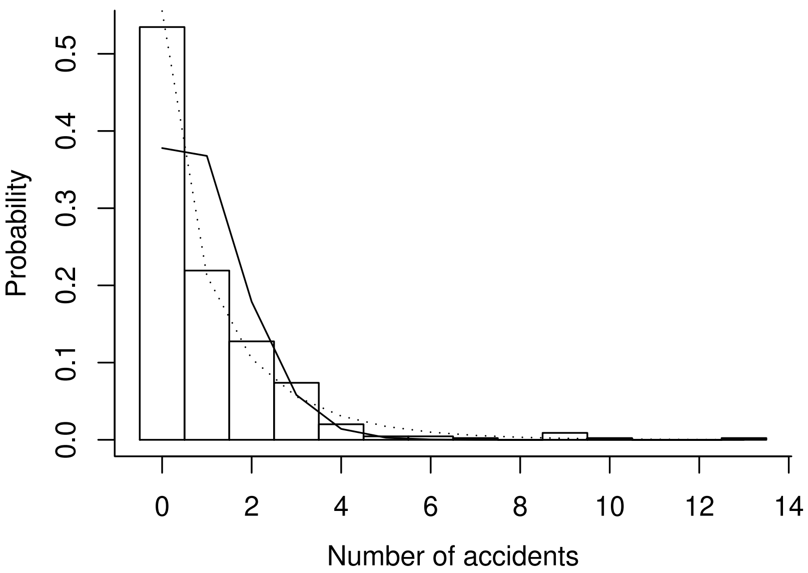

Figure 4.5: Histogram for the soap factory accident data, with fitted Poisson (solid) and negative binomial (dotted) distributions. (Irwin, 1975)

- A Poisson distribution with a different mean in each category gives a satisfactory fit. We would conclude that accidents happen randomly in each category, though with a different probability.
- the Poisson hypothesis is rejected in one or both categories, indicating further heterogeneity in the groups. A possible alternative would be the negative binomial. But note that we might end up with different distributions in the two categories, indicating that a different data generating mechanism is operating in each group.

Some other distribution (than the Poisson and the negative binomial) could also be used.

Question (6)

The table below gives the numbers of deaths by horse kicks in 10 Prussian army corps over a 20 year period, that is, for 200 corps–years (Sokal and Rohlf, 1969, p. 94).

Deaths	Corps
0	109
1	65
2	22
3	3
4	1

(a) What is a reasonable model to describe the way in which these deaths might have occurred?

(b) Fit the model and explain your conclusions.

Answer

This is the classical data set for the Poisson distribution!

(a) The assumption is that soldiers are killed at random by such kicks. In fact, there are ten distinct army corps observed over 20 years. Thus, there may be differences in risk of such death among the corps and it may be changing over time.

(b) The mean of the Poisson distribution is estimated to be $\hat{\mu} = 0.61$. The AIC is 2.87 as compared to 10 for the saturated model. It is unlikely that another model would have an AIC much small than this. The fitted model is displayed in Figure **??**, showing the small residuals. It is also plotted in Figure **??**.

Although knowledge of the background makes this conclusion implausible, the Poisson distribution does fit well, indicating randomness of the deaths.

Question (7)

The table below gives the number of fire losses per year from 1950 to 1973 for the buildings in a major university (Aiuppa, 1988).

Table 4.8: Frequency of deaths by horse kicks in the Prussian army with fitted Poisson distribution and standardised residuals.

Deaths	Years	Multinomial	Poisson	Residual
0	109	0.545	0.543	0.032
1	65	0.325	0.331	−0.158
2	22	0.110	0.101	0.396
3	3	0.015	0.021	−0.548
4	1	0.005	0.003	0.471

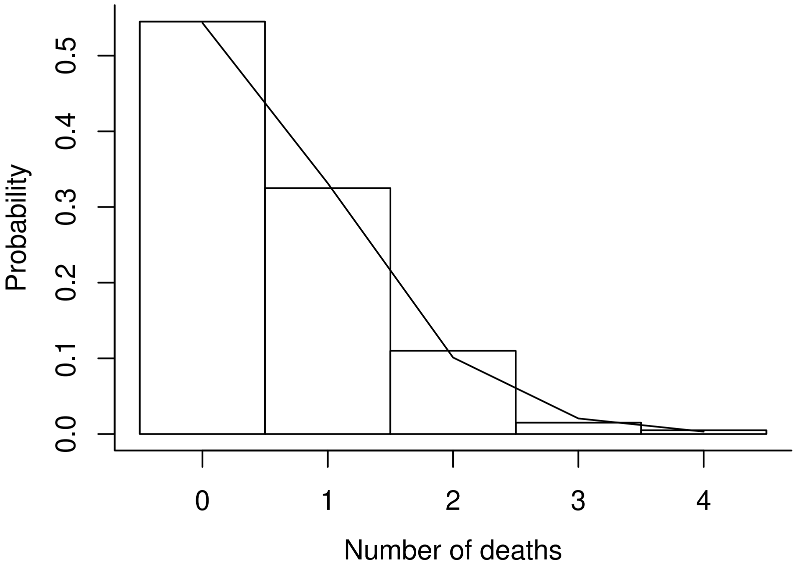

Figure 4.6: Histogram for the horse kicks data, with fitted Poisson distribution.

Table 4.9: Frequency of fire losses in a major university, with fitted Poisson distribution and standardised residuals. (Aiuppa, 1988)

Losses	Years	Multinomial	Poisson	Residual
0	0	0.000	0.031	−0.842
1	3	0.130	0.107	0.338
2	7	0.304	0.187	1.306
3	2	0.087	0.216	−1.335
4	5	0.217	0.188	0.323
5	1	0.043	0.131	−1.159
6	3	0.130	0.076	0.949
7	2	0.087	0.038	1.216

Number of losses	Number of years
0	0
1	3
2	7
3	2
4	5
5	1
6	3
7	2

(a) Choose an appropriate probability distribution and fit it.

(b) Calculate the deviance and check the residuals.

(c) Discuss how well the model fits.

(d) In fact, the number of buildings at the university evolved over the years concerned, from 273 in 1950 to 312 in 1962. If such data were available for each year, discuss how this could be taken into account.

Answer

(a) We may ask if fires are occurring at random in the university and hence fit a Poisson distribution. The maximum likelihood estimate of the mean is $\hat{\mu} = 3.48$. The results are given in Table **??**. The model is plotted in Figure **??**. We see that the model smooths the data a great deal.

(b) The deviance is 10.19. The residuals in Table **??** show no particular pattern and none are very large in spite of the large amount of smoothing.

(c) The AIC is 12.19. This compares with an AIC of 14 for the multinomial model, indicating that the Poisson model does not fit too badly. Fires might be occurring at random over the years. However, there are very few observations, so that this good fit is not very strong evidence. We really require more observations; calculating a sample size before the study will generally not be of too much use for fires in one university.

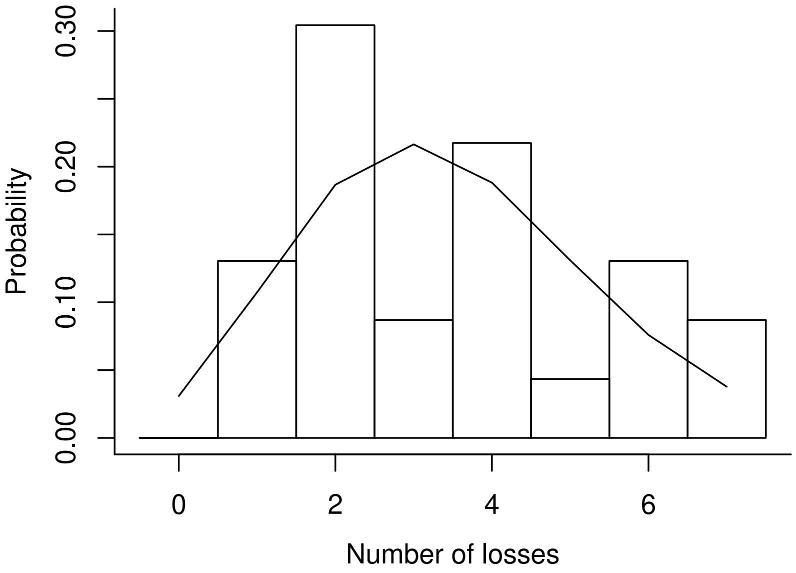

Figure 4.7: Histogram and fitted Poisson distribution for fire losses in a university. (Aiuppa, 1988)

Other distributions will probably also fit well, as can be seen by the difference between the histogram and fitted distribution in Figure **??**. The estimated mean and variance are almost identical: 3.49 and 3.55.

(c) In fact, we should be looking at the ratio of fires to buildings. If we were to use a log linear regression model in $\log(\mu_i)$ for the number of fires each year (say b_i), we could make it depend on the logarithm of the number of buildings, as a constant, without an unknown regression parameter, β, multiplying it:

$$\log(\mu_i) = \beta_0 + \log(b_i)$$

This is known as an offset.

Question (8)

Let us reconsider the consumer purchasing data of Exercise (**??**.**??**).

(a) Choose an appropriate probability distribution and fit it.

(b) Calculate the AIC and check the residuals.

(c) Discuss how well the model fits.

Answer

(a) For the purchase of consumer goods, we would suspect that the habits of all people are not homogeneous so that the Poisson hypothesis of random buying would not be applicable. The estimated means and variances are, respectively, $\bar{y}_\bullet = 0.64$, $s^2 = 4.50$ and $\bar{y}_\bullet = 0.68$, $s^2 = 6.87$ (coefficients of dispersion, 7.0 and 10.1), confirming this suspicion. Thus, we can try fitting a negative binomial distribution. The results are given in Table **??**, with $\hat{\nu}_1 \doteq 0.141$ and $\hat{\gamma} \doteq 0.105$, and Table **??**, with $\hat{\nu}_1 \doteq 0.098$ and $\hat{\gamma} \doteq 0.074$.

(b) The deviances and AICs are, respectively, 36.25 (40.25) and 23.01 (27.01), as compared to AICs of 54 and 42 for the multinomial, indicating a reasonable fit. The residuals show no systematic pattern, although there might be some concern with zero and small numbers of purchases, the pattern being reversed between the two tables.

(c) The Poisson and negative binomial models are plotted in Figures **??** and **??**. The latter appear to fit well, even surprisingly well, given the large numbers of zero purchases, although the problems with zero and one purchases are visible.

Question (9)

During a cholera epidemic in India, the number of cases in each house was recorded (Dahiya and Gross, 1973):

Cases	Houses
0	168
1	32
2	16
3	6
4	1

Table 4.10: Frequency of purchase of the first consumer good, with fitted negative binomial distribution and standardised residuals. (Chatfield, Ehrenberg, and Goodhardt, 1966)

Units	Households	Multinomial	Negative binomial	Residual
0	1612	0.8060	0.8148	−0.438
1	164	0.0820	0.0732	1.454
2	71	0.0355	0.0347	0.187
3	47	0.0235	0.0209	0.799
4	28	0.0140	0.0139	0.022
5	17	0.0085	0.0098	−0.599
6	12	0.0060	0.0072	−0.623
7	12	0.0060	0.0054	0.380
8	5	0.0025	0.0041	−1.118
9	7	0.0035	0.0032	0.261
10	6	0.0030	0.0025	0.467
11	3	0.0015	0.0020	−0.461
12	3	0.0015	0.0016	−0.062
13	5	0.0025	0.0012	1.595
14	0	0.0000	0.0010	−1.414
15	0	0.0000	0.0008	−1.270
16	0	0.0000	0.0007	−1.144
17	2	0.0010	0.0005	0.907
18	0	0.0000	0.0004	−0.932
19	0	0.0000	0.0004	−0.843
20	1	0.0005	0.0003	0.546
21	0	0.0000	0.0002	−0.692
22	2	0.0010	0.0002	2.555
23	0	0.0000	0.0002	−0.571
24	0	0.0000	0.0001	−0.519
25	1	0.0005	0.0001	1.645
26	2	0.0010	0.0001	4.221

Table 4.11: Frequency of purchase of the second consumer good, with fitted negative binomial distribution and standardised residuals. (Chatfield, Ehrenberg, and Goodhardt, 1966)

Units	Households	Multinomial	Negative binomial	Residual
0	1498	0.8580	0.8429	0.688
1	81	0.0464	0.0560	−1.700
2	47	0.0269	0.0271	−0.050
3	25	0.0143	0.0169	−0.830
4	16	0.0092	0.0117	−0.983
5	17	0.0097	0.0086	0.511
6	6	0.0034	0.0066	−1.611
7	10	0.0057	0.0051	0.348
8	3	0.0017	0.0041	−1.550
9	3	0.0017	0.0033	−1.156
10	6	0.0034	0.0027	0.586
11	4	0.0023	0.0022	0.049
12	4	0.0023	0.0019	0.418
13	3	0.0017	0.0016	0.170
14	2	0.0011	0.0013	−0.192
15	2	0.0011	0.0011	0.045
16	3	0.0017	0.0009	1.056
17	1	0.0006	0.0008	−0.340
18	0	0.0000	0.0007	−1.095
19	2	0.0011	0.0006	0.957
20	1	0.0006	0.0005	0.122
21	12	0.0069	0.0034	2.492

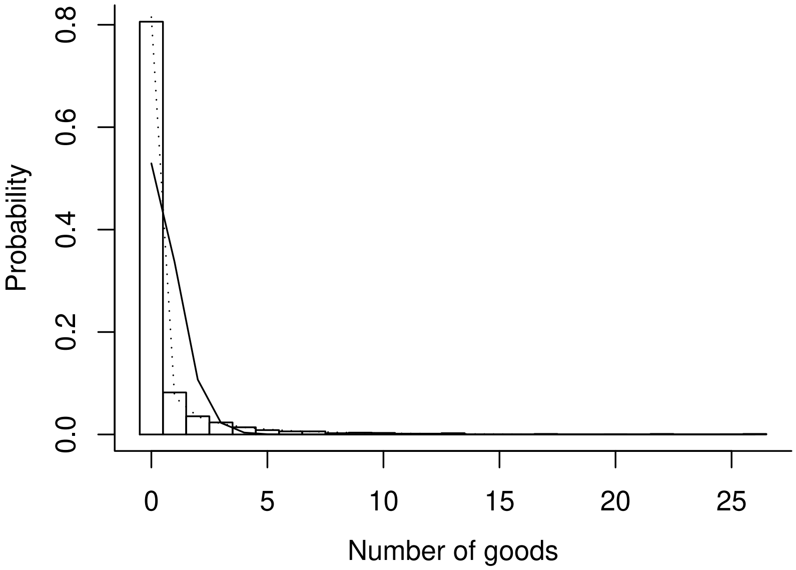

Figure 4.8: Histogram for purchases of the first consumer good, with fitted Poisson (solid) and negative binomial (dotted) distributions. (Chatfield, Ehrenberg, and Goodhardt, 1966)

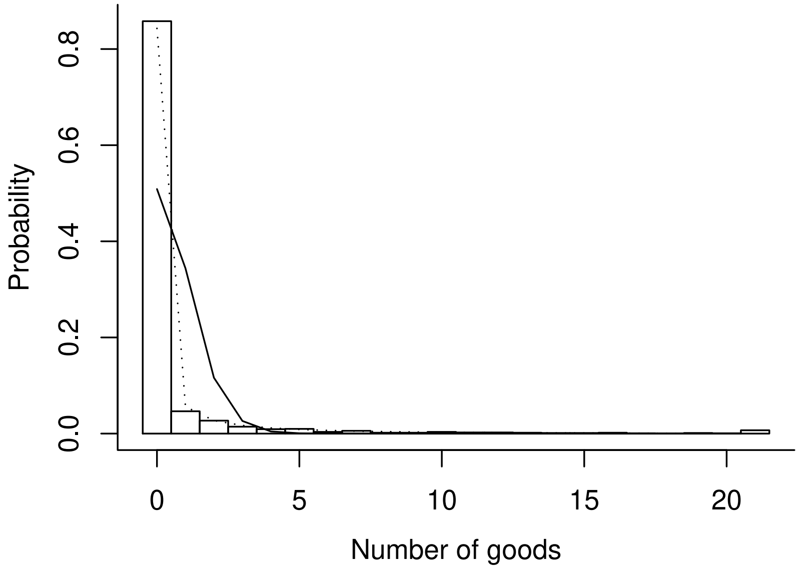

Figure 4.9: Histogram for purchases of the second consumer good, with fitted Poisson (solid) and negative binomial (dotted) distributions. (Chatfield, Ehrenberg, and Goodhardt, 1966)

Table 4.12: Frequency of cholera cases with fitted Poisson distribution and standardised residuals.

		Multinomial	Poisson		Negative binomial	
Cases	Houses		Fitted	Residual	Fitted	Residual
0	168	0.753	0.680	1.328	0.749	0.081
1	32	0.143	0.262	−3.463	0.167	−0.852
2	16	0.072	0.051	1.407	0.054	1.156
3	6	0.027	0.007	3.779	0.019	0.838
4	1	0.004	0.001	2.301	0.007	−0.468

(a) Fit an appropriate distribution to these data and draw your conclusions.

(b) Some of the houses which registered no cases were probably already infected. Fit a model without using this category and use it to predict the number of such houses among the 168.

Answer

This is data set looks similar to the Prussian horse kicks above. However, the results are not!

(a) A simple assumption might be that cholera cases were distributed at random in the population. However, this certainly should not be the case: cholera often results from polluted water, creating clumping of cases in certain areas.

(b) The mean of the Poisson distribution is estimated to be $\hat{\mu} = 0.39$. However, the AIC is 30.0 as compared to 10 for the saturated model. The fitted model is displayed in Figure **??** and plotted in Figure **??**.

The ratio of the variance to the mean is 1.54, perhaps indicating a little overdispersion. This is confirmed by fitting the negative binomial distribution, with an AIC of 8.9, a satisfactory fit. The results for this distribution are also shown in the table and graph. Apparently, there is clumping of the disease, perhaps among neighbours with the same water source.

Question (10)

In Table **??**, I studied the distribution by sex in families of 12 children in Saxony. The following table (Fisher, 1958, p. 67) gives the results from the same study for families of 8 children.

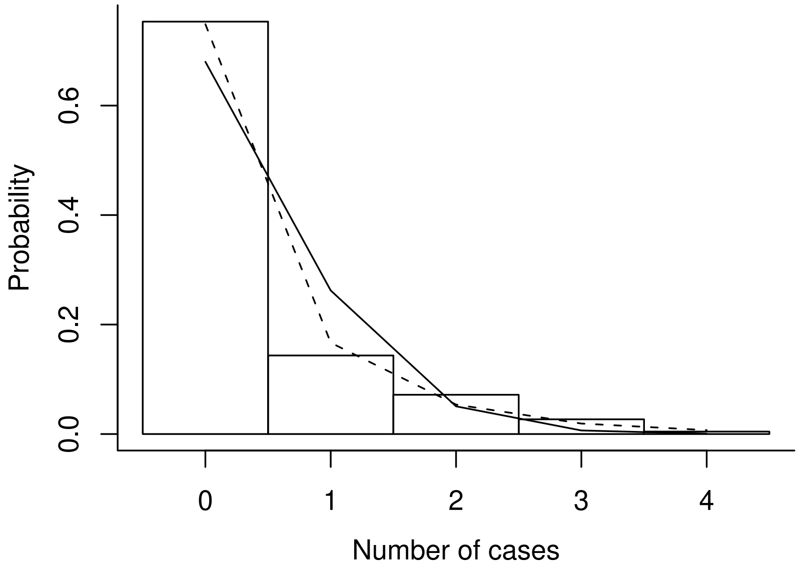

Figure 4.10: Histogram for the cholera data, with fitted Poisson (solid) and negative binomial (dashed) distributions.

Table 4.13: Frequency of male children in 6115 families of size 12 in Saxony, with fitted binomial distribution and standardised residuals. (Sokal and Rohlf, 1969, p. 80)

Males	Families	Multinomial	Binomial	Residual
0	3	0.000	0.000	2.140
1	24	0.004	0.002	3.423
2	104	0.017	0.012	3.793
3	286	0.047	0.042	1.708
4	670	0.110	0.103	1.676
5	1033	0.169	0.177	−1.591
6	1343	0.220	0.224	−0.658
7	1112	0.182	0.207	−4.323
8	829	0.136	0.140	−0.864
9	478	0.078	0.067	3.361
10	181	0.030	0.023	4.181
11	45	0.007	0.004	3.706
12	7	0.001	0.000	3.038

Table 4.14: Frequency of male children in 53 680 families of size eight in Saxony, with fitted binomial and beta binomial distributions and standardised residuals. (Fisher, 1958, p. 67)

		Multinomial	Binomial		Beta binomial	
Males	Families		Fitted	Residual	Fitted	Residual
0	215	0.004	0.003	3.873	0.004	1.779
1	1485	0.028	0.026	2.225	0.028	−0.586
2	5331	0.099	0.097	1.779	0.099	0.105
3	10649	0.198	0.206	−3.671	0.204	−2.861
4	14959	0.279	0.272	2.740	0.267	5.117
5	11929	0.222	0.231	−4.317	0.228	−3.028
6	6678	0.124	0.123	1.205	0.124	−0.026
7	2092	0.039	0.037	2.200	0.039	−0.599
8	342	0.006	0.005	4.780	0.006	2.428

Males	Families
0	215
1	1485
2	5331
3	10649
4	14959
5	11929
6	6678
7	2092
8	342

(a) Is a binomial distribution suitable in this case?

(b) Do the residuals indicate the same sort of departures from this model as for families of 12 children?

(c) What conclusions can be drawn from the analysis of the two data sets?

Answer

(a) The results for the fit of the binomial distribution are given in Table **??** and plotted in Figure **??**. The probability of a male is estimated to be $\hat{\nu}_1 = 0.515$. The deviance is 88.7 and the AIC 90.7, as compared to 16 for the multinomial distribution. (b) The mean and variance are estimated as $\bar{y}_{\bullet} = 4.12$ and $s^2 = 2.07$, whereas the theoretical variance is $8\hat{\nu}_1(1-\hat{\nu}_1) = 2.00$ which is close. Although there are two large negative residuals near the centre, there is no systematic pattern like that for families of 12 children. Thus, although the deviance is large, there appears to be no evidence here of overdispersion. Nevertheless, the beta binomial distribution does fit better, with an AIC of 56.7.

However, there are almost ten times as many families as for the table with 12 children. As we have seen from sample size calculations, large samples allow us to detect

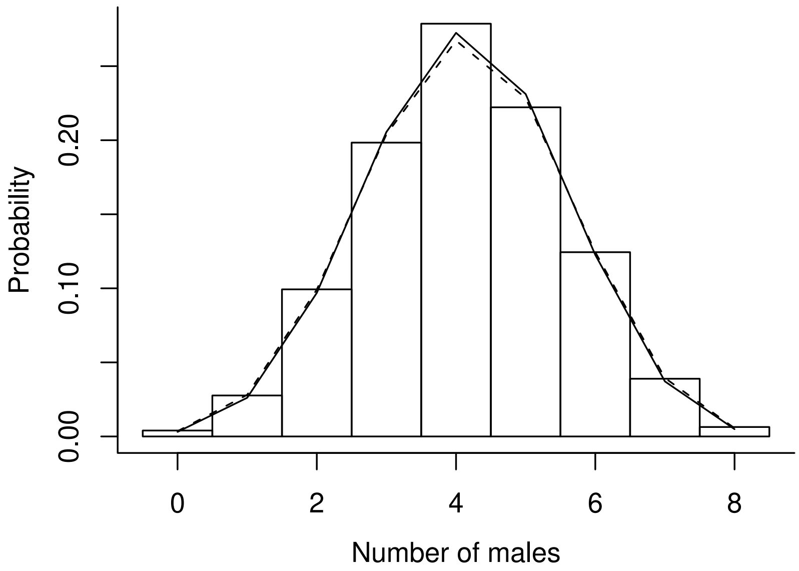

Figure 4.11: Histogram and fitted binomial (solid) and beta binomial (dashed) distributions for the male children in 53 680 families of size eight in Saxony. (Fisher, 1958, p. 67).

small departures from a model. That is probably what is happening here. If we had a sample one tenth this large with the same observed relative frequencies, the deviance would only be 8.87.

A more reasonable procedure, when the sample size is too large, is to increase the factor by which the number of parameters is multiplied in the AIC. For example, if we take ten for each parameter instead of two, the binomial AIC is 98.7 and the beta binomial 72.7 as compared to 80 for the multinomial distribution, even the former much closer. This is called the 'smoothing factor' because larger values yield a simpler smoother model. Note that the size of this factor should be decided before obtaining the data.

We may conclude that these models fits reasonably well, given the very large sample size.

(c) If families of 12 are overdispersed, but not those of 8, we may conclude that the former may arise from some special characteristics of those families. One hypothesis might be that parents with a large number of children of one sex continue having children in the hope of having at least one of the opposite sex. This would explain the over-representation of families with mostly males or mostly females in the table with 12 children.

Question (11)

The number of children ever born to a sample of mothers over 40 years of age was collected by the East African Medical Survey in the Kwimba district of Tanganyika (Brass, 1959):

Children	1	2	3	4	5	6	7	8	9	10	11	12
Mothers	49	56	73	41	43	23	18	18	7	7	3	2

(a) List the ways in which these data are different than those in Table **??** and in Exercise (**??**.**??**)

(b) Try to fit a suitable distribution for these data. (A somewhat similar data set, the postal survey, was given as an example in Section **??**. Recall also the discussion of truncated distributions in Section **??**.)

Answer

(a) The sibship data are a complete census of births in Saxony in the 19th century whereas this is a sample in Tanganyika in the 20th century. Those tables were each for one fixed size of family, whereas here all families are included. There, mothers were of all ages, whereas here they are all over 40 years old.

(b) One approach would be to follow up the hints in the question and try to fit truncated distributions. However, the sample only contains *mothers*. Hence, they must have at least one child. This is somewhat like the car occupant example in Section **??**. We can try using the number of children after the first as the response variable.

There is rather strong evidence for overdispersion: the ratio of variance to mean is about 2. The results for the Poisson and negative binomial distributions are shown

Table 4.15: Frequency of children in families in Tanganyika, with fitted Poisson and negative binomial distributions and standardised residuals

		Multinomial	Poisson		Negative binomial	
Children	Families		Fitted	Residual	Fitted	Residual
1	49	0.144	0.050	7.725	0.129	0.769
2	56	0.165	0.150	0.688	0.188	−0.981
3	73	0.215	0.225	−0.389	0.185	1.271
4	41	0.121	0.224	−4.030	0.153	−1.549
5	43	0.126	0.168	−1.850	0.115	0.605
6	23	0.068	0.100	−1.897	0.081	−0.880
7	18	0.053	0.050	0.245	0.055	−0.138
8	18	0.053	0.021	3.986	0.036	1.694
9	7	0.021	0.008	2.601	0.023	−0.245
10	7	0.021	0.003	6.420	0.014	1.021
11	3	0.009	0.001	5.256	0.009	0.051
12	2	0.006	0.000	7.113	0.005	0.185

in Table **??** and Figure **??**. The AICs are 120.3 for the Poisson and 19.4 for the negative binomial, as compared to 24 for the saturated model. The power parameter of the negative binomial is $\hat{\gamma} = 2.81$, considerably different than unity for the geometric distribution.

There are more families with three children than predicted by the negative binomial distribution and fewer with two and four. It might be interesting to investigate this further.

Question (12)

For the divorce data of Table **??**,

(a) How well does the inverse Gauss distribution fit to these data as compared with those tried in the text?

(b) What is a possible interpretation of this model in this context?

(c) Does any transformation of the data yield a reasonable model for these data?

Answer

(a) The inverse Gauss distribution fits much more poorly than either the gamma or Weibull, with an AIC of 157.8, as compared respectively to 51.8 and 86.1. The results are shown in Table **??** and Figure **??**. From the latter, we can see that the distribution peaks too soon.

(b) One interpretation of marriage might be as a series of changing tensions between the two spouses that might eventually lead to divorce. These data do not support the idea of such a theory with the tensions follow a random walk.

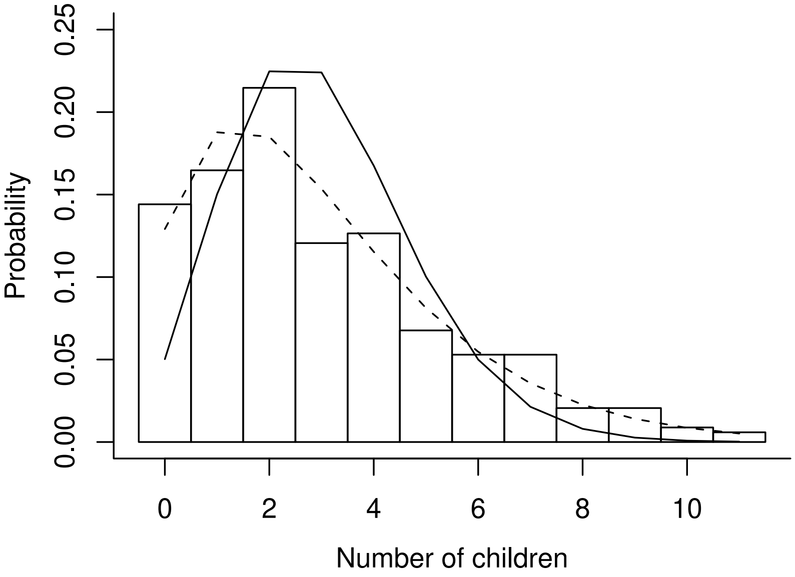

Figure 4.12: Histogram and fitted Poisson (solid) and negative binomial (dashed) distributions for children (excluding the first) in Tanganyika.

Table 4.16: Length of marriage (years) before divorce in Liège, 1984, with the fitted Weibull distribution. (Lindsey, 1992, pp. 14–15)

Years	Divorces	Mult.	Weibull	Years	Divorces	Mult.	Weibull
1	3	0.002	0.019	27	14	0.008	0.012
2	18	0.011	0.028	28	17	0.010	0.011
3	59	0.035	0.035	29	12	0.007	0.010
4	87	0.051	0.040	30	17	0.010	0.008
5	82	0.048	0.044	31	10	0.006	0.007
6	90	0.053	0.047	32	11	0.006	0.006
7	91	0.054	0.049	33	13	0.008	0.005
8	109	0.064	0.050	34	7	0.004	0.005
9	94	0.055	0.050	35	9	0.005	0.004
10	83	0.049	0.049	36	9	0.005	0.003
11	101	0.059	0.048	37	9	0.005	0.003
12	91	0.054	0.046	38	10	0.006	0.002
13	94	0.055	0.044	39	5	0.003	0.002
14	63	0.037	0.042	40	3	0.002	0.002
15	68	0.040	0.040	41	3	0.002	0.001
16	56	0.033	0.037	42	4	0.002	0.001
17	62	0.036	0.035	43	6	0.004	0.001
18	40	0.024	0.032	44	0	0.000	0.001
19	43	0.025	0.030	45	0	0.000	0.001
20	41	0.024	0.027	46	1	0.001	0.001
21	28	0.016	0.025	47	0	0.000	0.000
22	24	0.014	0.022	48	2	0.001	0.000
23	39	0.023	0.020	49	0	0.000	0.000
24	34	0.020	0.018	50	0	0.000	0.000
25	14	0.008	0.016	51	0	0.000	0.000
26	22	0.013	0.014	52	1	0.001	0.000

Table 4.17: Length of marriage (years) before divorce in Liège, 1984, with the fitted inverse Gauss distribution. (Lindsey, 1992, pp. 14–15)

Years	Divorces	Mult.	Weibull	Years	Divorces	Mult.	Weibull
1	3	0.002	0.000	27	14	0.008	0.009
2	18	0.011	0.007	28	17	0.010	0.008
3	59	0.035	0.028	29	12	0.007	0.008
4	87	0.051	0.050	30	17	0.010	0.007
5	82	0.048	0.063	31	10	0.006	0.006
6	90	0.053	0.069	32	11	0.006	0.006
7	91	0.054	0.070	33	13	0.008	0.005
8	109	0.064	0.067	34	7	0.004	0.005
9	94	0.055	0.063	35	9	0.005	0.004
10	83	0.049	0.058	36	9	0.005	0.004
11	101	0.059	0.053	37	9	0.005	0.003
12	91	0.054	0.048	38	10	0.006	0.003
13	94	0.055	0.043	39	5	0.003	0.003
14	63	0.037	0.039	40	3	0.002	0.003
15	68	0.040	0.035	41	3	0.002	0.002
16	56	0.033	0.031	42	4	0.002	0.002
17	62	0.036	0.028	43	6	0.004	0.002
18	40	0.024	0.025	44	0	0.000	0.002
19	43	0.025	0.022	45	0	0.000	0.002
20	41	0.024	0.020	46	1	0.001	0.001
21	28	0.016	0.018	47	0	0.000	0.001
22	24	0.014	0.016	48	2	0.001	0.001
23	39	0.023	0.014	49	0	0.000	0.001
24	34	0.020	0.013	50	0	0.000	0.001
25	14	0.008	0.012	51	0	0.000	0.001
26	22	0.013	0.010	52	1	0.001	0.001

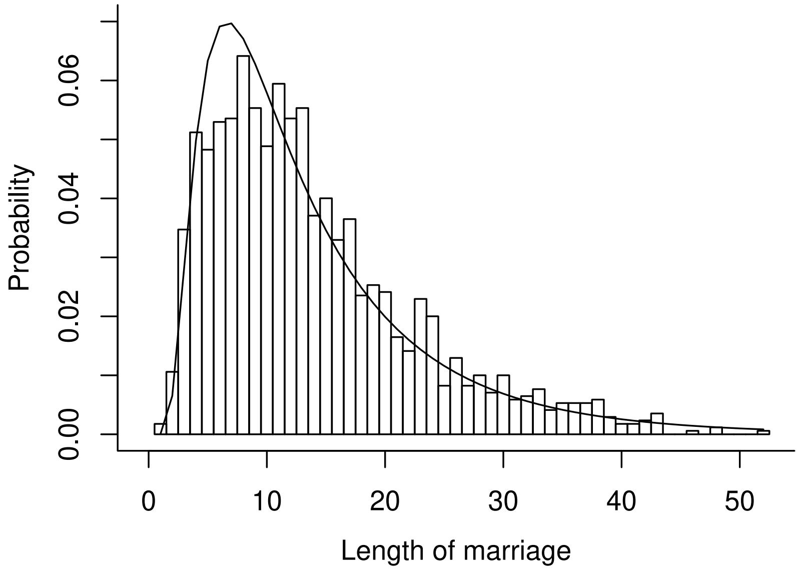

Figure 4.13: Histogram and fitted inverse Gauss distribution for the divorce data.

(c) The log normal distribution fits almost as badly as the inverse Gauss (AIC 128.2). Other more *ad hoc* transformations might also be tried.

Question (13)

The table below shows the duration of strikes in the U.K. which began in 1965, as recorded by the Ministry of Labour (Lancaster, 1972). Those given lasted more than one day and involved at least 10 people; they are for metal manufacturing and for all industries except transport and electrical machinery. One day strikes have not been included because they often are of a different nature, being a token stoppage appearing as a demonstration or threat. In the period considered, the majority of strikes in the U.K. were not claims for wage increases, but about questions of discipline, hours of work, sympathy, union recognition, and so on.

	Number of strikes			Number of strikes	
Duration	Metal	All	Duration	Metal	All
2	43	203	10	3	23
3	37	149	11–15	16	61
4	21	100	16–20	4	27
5	19	71	21–25	4	17
6	11	49	26–30	3	16
7	8	33	31–40	3	16
8	8	29	41–50	5	12
9	9	26	> 50	4	8

(a) Choose appropriate probability distributions and fit them to each data set.

(b) Compare the results.

(c) Calculate the deviances and check the residuals.

(d) Discuss how well the models fit to each and what can be learnt about the process by which a strike comes to a conclusion.

Answer

(a) From the description of the reasons for the strikes, we might think that the inverse Gauss distribution would be appropriate, as a model for changing satisfaction. However, from the histograms in Figures **??** and **??** for metal and all industries, respectively, we might be more inclined to choose the exponential distribution. We shall try both. One problem with these data is that we do not have information about strikes lasting only one day. The data are truncated on the left. We shall have to ignore this here but, depending on whether there were a lot of one day strikes, compared with the other short durations, this could greatly influence the results.

Note also that we could define the response in another way and study the number of supplementary days as compared to one day strikes instead of the strike length; this would eliminate the left truncation problem which, strictly speaking, would require us to renormalised the probabilities (as the zero frequency corresponding to one day strike is a structural zero).

(b) (c) The deviances (AIC) are 119.33 (121.33) and 511.49 (513.49) for the exponential and 43.57 (47.57) and 178.19 (182.19) for the inverse Gauss. The AIC for the multinomial is 30 so that none fit well, although the inverse Gauss is considerably better than the exponential. From Figures **??** and **??**, we see that the exponential does not drop fast enough. The Pareto distribution may be an alternative. It has deviances and AICs of 38.77 (42.77) and 146.49 (150.49). This is somewhat better but does not have a useful interpretation.

The estimated parameters are $\hat{\mu} = 8.71$ and 7.84 days, respectively, for the exponential distribution, $\hat{\mu} = 8.71$ and 7.84, $\hat{\sigma}^2 = 0.138$ and 0.135 for the inverse Gauss, and $\hat{\delta} = 1.5$, $\hat{\alpha} = 0.792$ and 0.826 for the Pareto.

The results for this inverse Gauss model are given in Tables **??** and **??**. We see that the model under-predicts the number of short strikes and over-predicts medium ones. The graphs for all three models are plotted in Figures **??** and **??**.

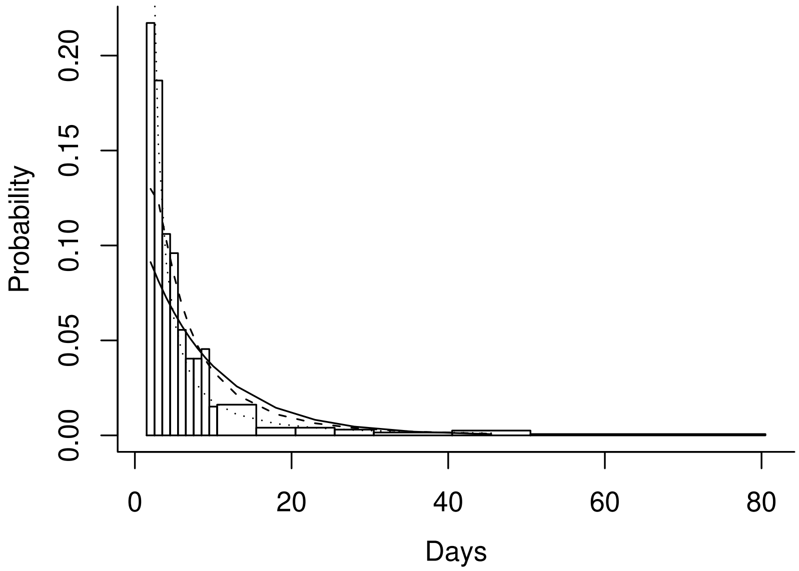

Figure 4.14: Histogram for the strikes in the metal industries, with fitted exponential (solid), Pareto (dashed), and inverse Gauss (dotted) distributions. (Lancaster, 1972)

Table 4.18: Frequency of different lengths of strikes in the metal industries, with fitted inverse Gauss distribution and standardised residuals. (Lancaster, 1972)

Duration	Strikes	Multinomial	Inverse Gauss	Residual
2	43	0.217	0.130	3.412
3	37	0.187	0.123	2.563
4	21	0.106	0.103	0.138
5	19	0.096	0.084	0.575
6	11	0.056	0.069	−0.711
7	8	0.040	0.057	−0.965
8	8	0.040	0.047	−0.442
9	9	0.045	0.040	0.407
10	3	0.015	0.034	−1.418
11–15	16	0.081	0.107	−1.123
16–20	4	0.020	0.056	−2.124
21–25	4	0.020	0.032	−0.919
26–30	3	0.015	0.019	−0.414
31–40	3	0.015	0.019	−0.427
41–50	5	0.025	0.008	2.561
>50	4	0.020	0.072	−2.728

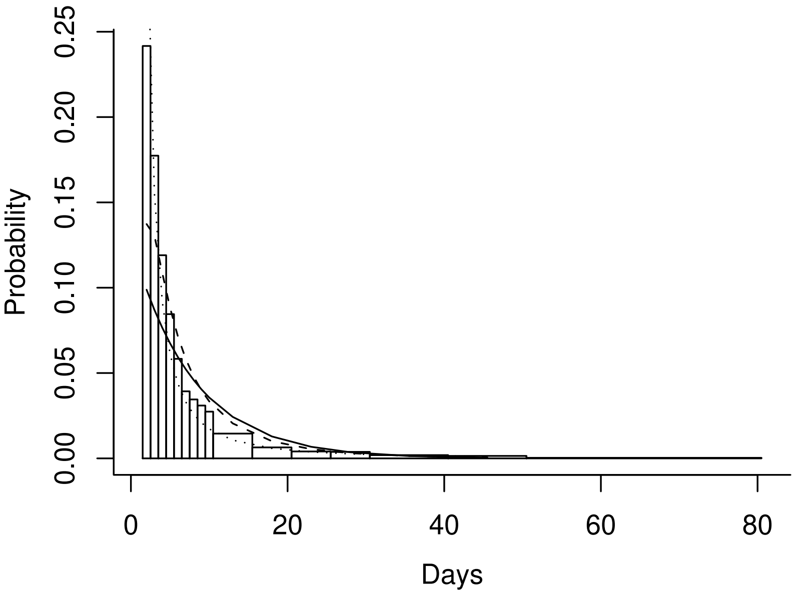

Figure 4.15: Histogram for the strikes in all industries, with fitted exponential (solid), Pareto (dashed), and inverse Gauss (dotted) distributions. (Lancaster, 1972)

Table 4.19: Frequency of different lengths of strikes in all industries, with fitted inverse Gauss distribution and standardised residuals. (Lancaster, 1972)

Duration	Strikes	Multinomial	Inverse Gauss	Residual
2	203	0.242	0.137	8.164
3	149	0.177	0.131	3.757
4	100	0.119	0.109	0.907
5	71	0.085	0.088	−0.356
6	49	0.058	0.071	−1.423
7	33	0.039	0.058	−2.283
8	29	0.035	0.048	−1.784
9	26	0.031	0.040	−1.296
10	23	0.027	0.033	−0.955
11–15	61	0.073	0.102	−2.699
16–20	27	0.032	0.050	−2.349
21–25	17	0.020	0.027	−1.184
26–30	16	0.019	0.015	0.886
31–40	16	0.019	0.014	1.239
41–50	12	0.014	0.005	3.508
>50	8	0.010	0.070	−6.616

(d) If the inverse Gauss distribution had been satisfactory, we might have thought that the strikes tended to end either when the to and fro of negotiations resulted in agreement or when the inconvenience to the strikers rose to a level that brought them back to work. However, the poor fit seems to exclude such explanations.

On the other hand, different results might be obtained if we could distinguish different types of strikes and study them separately, even separating long and short strikes.

Question (14)

A survey was made of women having a bachelor's but no higher degree and employed as mathematicians or statisticians. Monthly salaries (dollars) of these female mathematics graduates involved in non-supervisory positions are given below (Zelterman, 1987).

Monthly		Monthly		Monthly	
salary	No.	salary	No.	salary	No.
1051–1150	1	2151–2250	11	3251–3350	1
1151–1250	1	2251–2350	6	3351–3450	4
1251–1350	6	2351–2450	11	3451–3550	1
1351–1450	3	2451–2550	3	3551–3650	2
1451–1550	4	2551–2650	4	3651–3750	0
1551–1650	3	2651–2750	5	3751–3850	2
1651–1750	9	2751–2850	6	3851–3950	0
1751–1850	6	2851–2950	4	3951–4050	0
1851–1950	5	2951–3050	4	4051–4150	1
1951–2050	16	3051–3150	5		
2051–2150	4	3151–3250	1		

(a) Choose an appropriate probability distribution and fit it.

(b) Calculate the deviance and check the residuals.

(c) Discuss how well the model fits.

Answer

(a) Often, the normal distribution would automatically be used for such data. We fit it first and obtain a deviance of 46.65 with AIC 50.65, as compared to 60 for the multinomial. This appears to be a good fit. However, if we look at the plot, in Figure **??**, we may have the impression that this distribution is too wide, especially because of the large number of women with $2000 per month. Thus, we may want to try the logistic distribution, which, however, has a deviance of 50.85 and AIC 54.85. As seen in the plot, this is indeed narrower, but the fit is poorer.

Salaries are often skewed, with only a few high ones. We can check this for these data by fitting the log normal distribution, with deviance of 39.65 and AIC 43.65. This is considerably better than the previous models, although this may not be evident from the plot, also in Figure **??**. (The gamma and inverse Gauss distributions, with respective

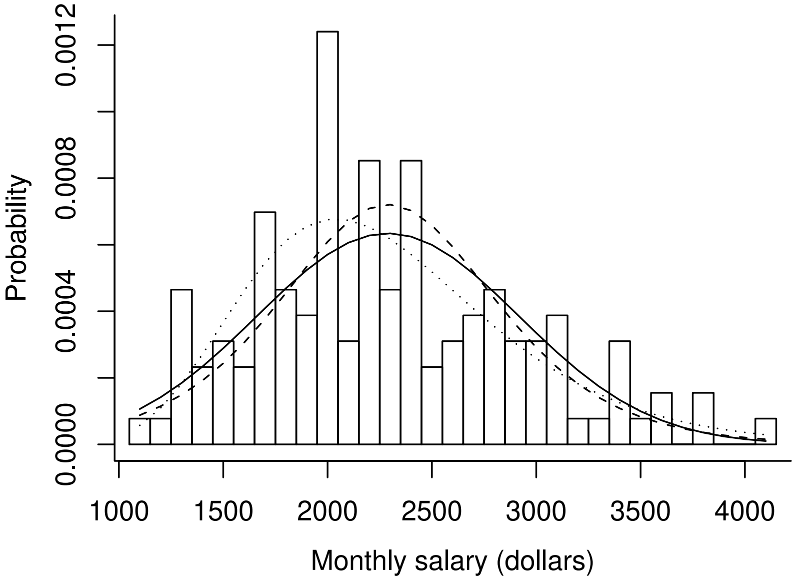

Figure 4.16: Histogram for women mathematicians' salaries, with fitted normal (solid), logistic (dashed), and log normal (dotted) distributions. (Zelterman, 1987)

AICs of 43.23 and 43.38, also fit equally as well as the log normal, illustrating that there may be no best model; this can be a useful exercise to convince students that there is no correct model.) The parameter estimates are $\hat{\mu} = 7.70$ and $\hat{\sigma}^2 = 0.077$. The former is the mean of the log salaries. From this, the mean salary is estimated to be 2290.54, from the formula near the bottom of page 117 in the text, which may be compared to the usual average, $\bar{y}_\bullet = 2289.15$.

(b) Because of the size of the table, we only provide the fitted values for the log normal distribution, in Table **??**. There appears to be no pattern in the residuals.

(c) We have very few observations here, so that our conclusions above, for example about skewness, must be provisional. It is interesting to note the high frequencies for the categories centred on \$2000, \$2200, and \$2400. If this tendency was confirmed in a larger sample, it would be difficult to find any distribution to fit the data adequately. As it is, the model smooths the data a great deal.

Question (15)

The following table gives the number of years since their first degree of the same sample of female mathematics graduates practising mathematics or statistics as described in Exercise (**??**.**??**) above. (Zelterman, 1987).

Years	Number	Years	Number	Years	Number
0	5	10	2	22–23	3
1	14	11	3	24–25	4
2	10	12	3	26–27	3
3	8	13	3	28–29	1
4	11	14	0	30–31	1
5	4	15	1	32–33	2
6	3	16	5	34–35	0
7	5	17	2	36–40	6
8	7	18–19	9		
9	5	20–21	9		

(a) Choose an appropriate probability distribution and fit it.

(b) Calculate the deviance and check the residuals.

(c) Discuss how well the model fits.

Answer

(a) These are duration data. Most distributions for such data cannot handle zero times, so we shall ignore them for this analysis. This can be justified in that these women do not yet have any experience and will be different than the others. The histogram, in Figure **??**, is more or less decreasing so that we may expect the exponential distribution to fit well. This might arise if the number of women graduating, and finding mathematics or statistics jobs, was a Poisson process over the years. The mean for this distribution is estimated to be $\hat{\mu} = 11.94$ years experience.

Table 4.20: Frequency distribution of women mathematicians' salaries, with fitted log normal distribution and standardised residuals. (Zelterman, 1987)

Salary	Women	Multinomial	Log normal	Residual
1051–1150	1	0.008	0.006	0.306
1151–1250	1	0.008	0.011	−0.344
1251–1350	6	0.047	0.018	2.389
1351–1450	3	0.023	0.027	−0.260
1451–1550	4	0.031	0.037	−0.337
1551–1650	3	0.023	0.046	−1.212
1651–1750	9	0.070	0.055	0.738
1751–1850	6	0.047	0.061	−0.674
1851–1950	5	0.039	0.066	−1.188
1951–2050	16	0.124	0.068	2.468
2051–2150	4	0.031	0.067	−1.591
2151–2250	11	0.085	0.065	0.889
2251–2350	6	0.047	0.062	−0.695
2351–2450	11	0.085	0.057	1.340
2451–2550	3	0.023	0.052	−1.425
2551–2650	4	0.031	0.046	−0.806
2651–2750	5	0.039	0.041	−0.110
2751–2850	6	0.047	0.035	0.672
2851–2950	4	0.031	0.030	0.040
2951–3050	4	0.031	0.026	0.365
3051–3150	5	0.039	0.022	1.306
3151–3250	1	0.008	0.018	−0.881
3251–3350	1	0.008	0.015	−0.682
3351–3450	4	0.031	0.012	1.881
3451–3550	1	0.008	0.010	−0.281
3551–3650	2	0.016	0.008	0.884
3651–3750	0	0.000	0.007	−0.938
3751–3850	2	0.016	0.006	1.525
3851–3950	0	0.000	0.004	−0.759
3951–4050	0	0.000	0.004	−0.681
4051–4150	1	0.008	0.003	1.030

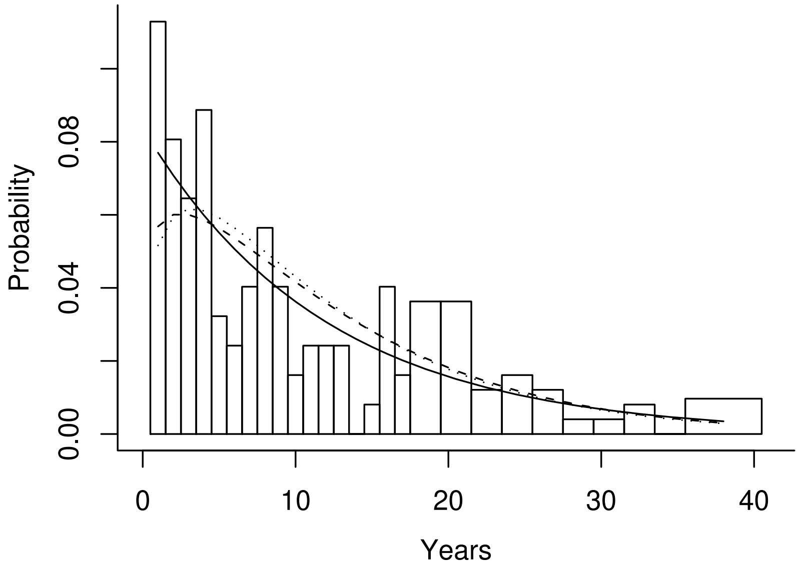

Figure 4.17: Histogram for women mathematicians' experience, with fitted exponential (solid), Weibull (dashed), and gamma (dotted) distributions. (Zelterman, 1987)

Note that, here, the choice of the y_is representing each category of the response is not as clear cut as one might expect. It was simply taken to be the recorded number of years of experience for the first categories and the average of the lower and upper bounds for the latter ones. However, the following argument should raise questions about this choice. Assume that mathematicians with, say, twenty months of experience were recorded as one year because the second year was not yet completed. Then, one could expect to have the category 'one year' stand for women with experience in between twelve and twenty four months. Hence, a reasonable choice for the corresponding y_i could be 1.5. But this argument assumes that graduation occurs randomly over the year, which is probably not the case in practice. Thus, the choice of y_i depends rather on the time at which the survey was performed. One could also argue that, if graduation occur at a fixed time in the year, then a discrete structure for the response should be assumed. Here, for simplicity, the first mentioned definition of y_i will be adopted. Note that our first argument would apply if the response was the number of years the questioned women have worked as a whole since they graduated, because unemployment can occur more or less randomly over a year. The midpoint for the category $(18,19)$ would then be 19.

This discussion really points out the care required to choose the y_i's and the lack of information on published data sets in statistics journals.

(b) This exponential distribution has deviance 56.13 and AIC 58.13. The multinomial AIC is 52, so that this is not a very good fit. The low numbers with 5–7 and 14–15 years experience would explain this poor fit. No other simple smooth distribution will correct this. The number of graduates over the years might be investigated to study this phenomenon further. The fitted model is described in Table **??**. The years that we thought to be under-represented above do not have exceptionally large residuals. On the other hand, the model under-estimates 18–21 and 36–40 years experience. This can also be seen in Figure **??**.

(c) The exponential distribution is a special case of both the gamma and Weibull distributions, so it may be useful to fit them to see if we can obtain any improvement. The gamma, with $\hat{\alpha} \doteq 1.37$ and $\hat{\mu} \doteq 11.94$, has an AIC of 59.74, whereas the Weibull, with $\hat{\alpha} \doteq 1.17$ and $\hat{\mu} \doteq 19.60$, has an AIC of 58.26. Thus, neither is better. This indicates that α can be unity in both of these distributions. Both are plotted in Figure **??**. Note how the curve rises to a maximum on the left for both distributions, in contrast to the exponential.

Question (16)

Let us look again at the event recall data of Exercise (**??**.**??**).

(a) What might be an appropriate probability distribution for these data?

(b) Fit the model.

(c) Calculate the AIC and check the residuals.

(d) Discuss how well the model fits.

Table 4.21: Frequency distribution of women mathematicians' experience, with fitted exponential distribution and standardised residuals. (Zelterman, 1987)

Years	Women	Multinomial	Exponential	Residual
0	5	–	–	–
1	14	0.113	0.077	1.441
2	10	0.081	0.071	0.411
3	8	0.065	0.065	−0.027
4	11	0.089	0.060	1.311
5	4	0.032	0.055	−1.083
6	3	0.024	0.051	−1.310
7	5	0.040	0.047	−0.324
8	7	0.056	0.043	0.732
9	5	0.040	0.039	0.051
10	2	0.016	0.036	−1.177
11	3	0.024	0.033	−0.557
12	3	0.024	0.031	−0.411
13	3	0.024	0.028	−0.265
14	0	0.000	0.026	−1.793
15	1	0.008	0.024	−1.138
16	5	0.040	0.022	1.383
17	2	0.016	0.020	−0.317
18–19	9	0.073	0.036	2.184
20–21	9	0.073	0.030	2.727
22–23	3	0.024	0.025	−0.088
24–25	4	0.032	0.022	0.814
26–27	3	0.024	0.018	0.494
28–29	1	0.008	0.015	−0.658
30–31	1	0.008	0.013	−0.484
32–33	2	0.016	0.011	0.542
34–35	0	0.000	0.009	−1.075
36–40	6	0.048	0.017	2.619

Table 4.22: Frequency of times of recall of stressful events, with fitted exponential distribution and standardised residuals. (Haberman, 1978, p. 3)

Months	Events	Multinomial	Exponential	Residual
1	15	0.102	0.119	−0.599
2	11	0.075	0.104	−1.093
3	14	0.095	0.091	0.186
4	17	0.116	0.079	1.577
5	5	0.034	0.069	−1.614
6	11	0.075	0.060	0.724
7	10	0.068	0.053	0.822
8	4	0.027	0.046	−1.053
9	8	0.054	0.040	0.877
10	10	0.068	0.035	2.154
11	7	0.048	0.030	1.196
12	9	0.061	0.027	2.582
13	11	0.075	0.023	4.119
14	3	0.020	0.020	0.018
15	6	0.041	0.018	2.119
16	1	0.007	0.015	−0.838
17	1	0.007	0.013	−0.692
18	4	0.027	0.012	1.739

Answer

(a) The exponential distribution appears to be appropriate for these data, both because stressful events should happen fairly randomly in time and because of the shape of the histogram plotted in Figure **??**.

(b) When we fit the model, we obtain the results shown in Table **??**. The mean length of recall is estimated to be $\hat{\mu} = 7.33$ months. There are some large residuals, with under-estimation in the middle range of times.

(c) We obtain a deviance of 80.58 and AIC of 82.58, as compared to 34 for the multinomial AIC. One problem is that events are truncated after 18 months. From the large number in the last month, we might think that certain people with older events classified them there. If we take the last category to mean ≥ 18, the deviance is reduced to 59.73 (61.73), still not a good fit.

(d) In Exercise **??**, we fitted a log linear regression to these data. This was equivalent to fitting a Poisson process. The rate or intensity of events per unit time for an individual is given by negative of β_1. The maximum likelihood estimate was -0.084, so that the intensity is estimated to be 0.084 events per month. (We do not use the approximate value calculated there, and plotted in Figure **??**, because it distorts the calculations to follow.) We know that the mean time between events, for an exponential distribution, is the reciprocal, i.e. 11.9 months. This is considerably longer than the estimate given above because fitting by log linear regression carries the assumption that observations were truncated at 18 months, more realistic than the above procedure. We can also calculate the deviance for this approach, obtaining 24.57, an acceptable

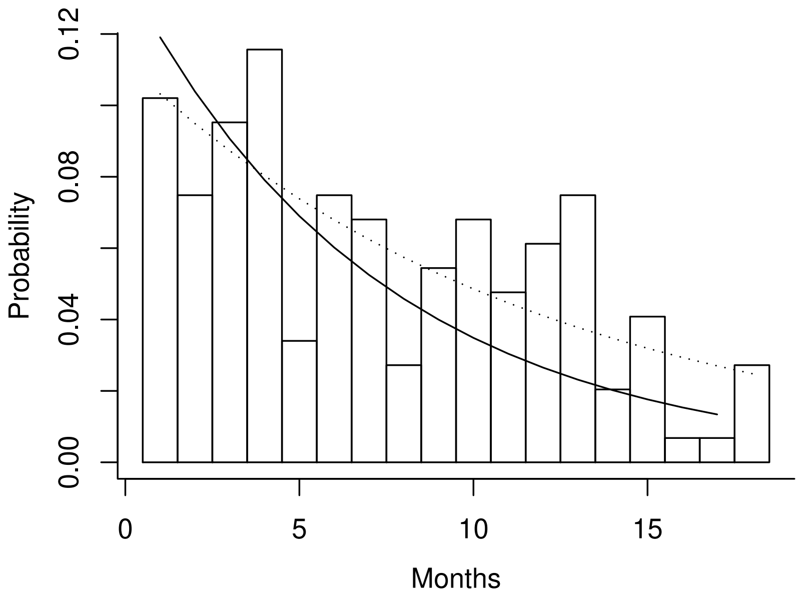

Figure 4.18: Histogram for the recall of stressful events, with fitted exponential (solid) and truncated exponential (dotted) distributions. (Haberman, 1978, p. 3)

model. This indicates that there should be a considerable number of people having had their last stressful event longer ago than 18 months. The models calculated by the two methods are displayed in Figure **??**. The graph confirms the superior fit of the truncated exponential distribution, supporting the idea that stressful events may be happening at random to individuals.

Notice the difference between censoring and truncation in this example. Censoring means that we have observed all events but the long times are recorded as 18 months. Truncation means that events over 18 months before are missing.

Question (17)

Employment durations of recruits to the British Post Office in the first quarter of 1973 (Burridge, 1981) were given in Table 4.6. In fact, there were two groups corresponding to different grades, as shown in the following table.

	Group			Group			Group	
Quarters	A	B	Quarters	A	B	Quarters	A	B
1	22	30	9	2	3	17	1	1
2	18	28	10	1	0	18	1	0
3	19	31	11	0	0	19	3	2
4	13	14	12	1	1	20	1	0
5	5	10	13	0	1	21	1	3
6	6	6	14	0	0	22	0	1
7	3	5	15	0	0	23	0	1
8	2	2	16	1	1	24	0	0

(a) Choose and fit a suitable probability distribution to each group.

(b) Compare the results and discuss how well each model fits.

(c) Have you found a continuous distribution that fits better than the geometric distribution?

(d) Calculate the AICs and check the residuals.

(e) Combine the data for the two groups and refit the model.

(f) Does it change very much from the two separate models?

(g) Because observations on the two groups are independent, the AICs for the two groups separately can be added together and compared to that for the model where the groups were combined. What can be said about the difference between the groups?

Answer

(a) If the recruits are leaving at random, we would expect an exponential distribution, whereas if they leave after reaching a certain level of dissatisfaction, we would rather look for an inverse Gauss distribution.

(b) The parameter estimates are $\hat{\mu} = 4.54$ and 4.31 for the exponential distribution; they are $\hat{\mu} = 4.54$, $\hat{\sigma}^2 = 0.221$ and $\hat{\mu} = 4.31$, $\hat{\sigma}^2 = 0.216$ for the inverse Gauss distribution. These estimates are very similar for the two types of recruits.

(c) The deviances (AICs) are 56.51 (58.51) and 85.07 (87.07) for the two groups for the exponential distribution and 24.79 (28.79) and 35.28 (39.28) for the inverse Gauss. By comparison, the multinomial AIC is 26, so that the inverse Gauss distribution is close to fitting well, especially for the first group. The log normal, gamma, and Weibull distributions fit much more poorly.

(d) When the data are combined, the deviances (AICs) are, respectively, 130.10 (132.10) and 48.61 (52.61) for the two distributions. Notice how they increase due to the larger sample available through combination.

(e) The parameter estimates are now $\hat{\mu} = 4.40$ for the exponential distribution and $\hat{\mu} = 4.40$, $\hat{\sigma}^2 = 0.218$ for the inverse Gauss distribution. As might be expected, they

Table 4.23: Frequency of employment times of recruits to the Post Office, with fitted inverse Gauss distribution and standardised residuals. (Burridge, 1981)

Quarters	Recruits	Multinomial	Inverse Gauss	Residual
1	52	0.2167	0.2172	−0.017
2	46	0.1917	0.2146	−0.768
3	50	0.2083	0.1521	2.234
4	27	0.1125	0.1063	0.297
5	15	0.0625	0.0758	−0.746
6	12	0.0500	0.0553	−0.348
7	8	0.0333	0.0412	−0.600
8	4	0.0167	0.0311	−1.274
9	5	0.0208	0.0240	−0.314
10	1	0.0042	0.0187	−1.643
11	0	0.0000	0.0147	−1.876
12	3	0.0083	0.0116	−0.475
13	1	0.0042	0.0093	−0.826
14	0	0.0000	0.0075	−1.341
15	0	0.0000	0.0061	−1.207
16	2	0.0083	0.0049	0.747
17	2	0.0083	0.0040	1.045
18	1	0.0042	0.0033	0.227
19	5	0.0208	0.0027	5.353
20	1	0.0042	0.0023	0.617
21	4	0.0167	0.0019	5.276
22	1	0.0042	0.0016	1.016
23	1	0.0042	0.0013	1.222
23	0	0.0000	0.0011	−0.513

are in between those for the two groups individually, and thus have not changed very much. The deviance to compare the separate models with the combined one is

$$-2\sum_{i}[(n_{1i}+n_{2i})\log(\tilde{\pi}_{\bullet i})-n_{1i}\log(\tilde{\pi}_{1i})-n_{2i}\log(\tilde{\pi}_{2i})]$$

where the indices indicate the two models. This gives a value of 0.18, indicating very little difference so that the two groups can be combined.

The fitted model for the inverse Gauss distribution is shown in Table **??**. This distribution, and the exponential, are plotted in Figure **??**. We see how the inverse Gauss distribution tries to account for the relatively equal numbers leaving in the first three quarters. This model over-estimates the shorter times (except for three and four quarters) and under estimates the longer ones.

We can conclude that the two groups are very similar and that recruits are not leaving randomly. There is some indication that a model of dissatisfaction or incompatibility describes the leaving process.

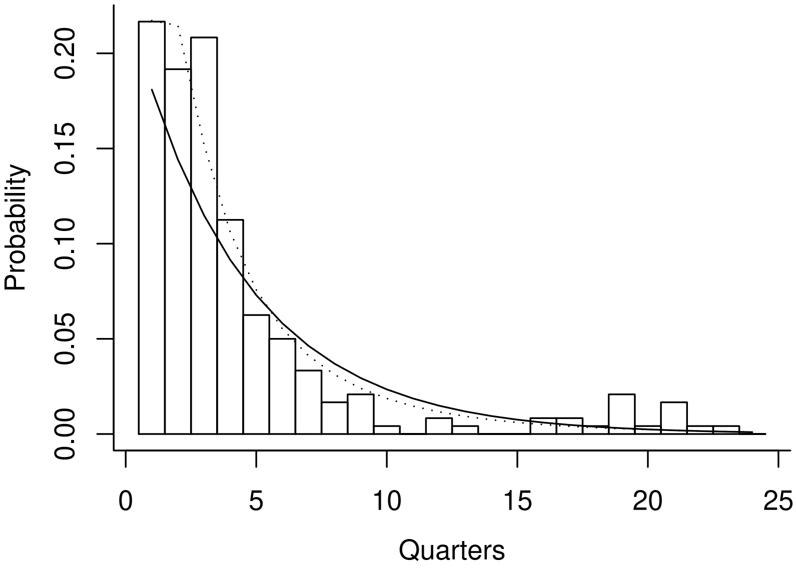

Figure 4.19: Histogram for employment times of recruits to the Post Office, with fitted exponential (solid) and inverse Gauss (dotted) distributions. (Burridge, 1981)

Question (18)

The inverse Gauss and log normal distributions are closely related to the normal distribution. The latter is in the location–scale family. Are the first two members of this family? Why?

Answer

Question (19)

For each distribution studied in this chapter,

(a) check whether it is a member of the exponential family,

(b) a member of the exponential dispersion family, and

(c) for the members of the exponential family, derive the sample size formula.

Answer

(a) Only the binomial, Poisson, and exponential distributions are members of the (one parameter) exponential family. The Poisson can be written

$$\log[f(y_i;\mu)] = -\mu + y_i \log(\mu) - \log(y_i!)$$

so that $\theta = \log(\mu)$, $a(\theta) = -\mu = -e^{\theta}$, and $b(y_i) = -\log(y_i!)$. Here, the canonical parameter is the logarithm of the mean, the basis of log linear models (just as the canonical parameter for the binomial, the logit, is the basis of logistic models).

The exponential distribution can be written

$$\log[f(y_i;\lambda)] = \log(\lambda) - \lambda y_i$$

so that the canonical parameter is $\theta = \lambda = 1/\mu$, with $a(\theta) = \log(\theta)$ and $b(y_i) = 0$.

(b) The normal, log normal, gamma, and inverse Gauss distributions are members of the exponential dispersion family. The log normal distribution has a form very similar to that for the normal given in the text. The gamma distribution can be written

$$\log[f(y_i;\mu,\alpha)] = \alpha[\log(\alpha) - \log(\mu)] - \log[\Gamma(\alpha)] + (\alpha-1)y_i - \alpha y_i/\mu$$

If we take $\theta = -1/\mu$ and $\sigma^2 = 1/\alpha$, we have $a(\theta) = -\log(\mu) = \log(-\theta)$ and $b(y_i,\sigma^2) = \alpha\log(\alpha) - \log[\Gamma(\alpha)] + (\alpha-1)y_i$.

For the inverse Gauss distribution,

$$\log[f(y_i;\mu,\sigma^2)] = -y_i/(2\sigma^2\mu^2) + 1/(\sigma^2\mu) - 1/(2y_i\sigma^2) - \log(2\pi y_i^3\sigma^2)/2$$

We can take $\theta = -1/\mu^2$ and σ^2, then $a(\theta) = 1/\mu = \sqrt{(}-1/\theta)$ and $b(y_i,\sigma^2) = -1/(2y_i\sigma^2) - \log(2\pi y_i^3\sigma^2)/2$.

(c) For the Poisson distribution, the sample size is calculated from

$$n_\bullet = \frac{D/2}{\bar{y}_\bullet[\log(\hat{\mu}) - \log(\mu_0)] + \mu_0 - \hat{\mu}}$$

whereas, for the exponential distribution, it is

$$n_\bullet = \frac{D/2}{\bar{y}_\bullet(\hat{\lambda} - \lambda_0) - \log(\lambda_0) + \log(\hat{\lambda})}$$

For a fixed value of σ^2, the sample size can also be calculated by this method for members of the exponential dispersion family. See Section **??** for the normal distribution.

Chapter 5

Normal regression and ANOVA

5.1 General regression models

Chapter **??** should have brought the students to the point where they should be eager to study more complex situations in which distributions vary with explanatory variables, i.e. combining Chapters **??** and **??**. Unfortunately, this cannot realistically be done in an introductory course, although it was partially tackled in Section **??** where we studied differences in the distribution of ages of school children in Bombay and in Exercise (**??**.**??**) on employment time in the British Post Office (where the explanatory variable was the grade of the employee); I tell them that they must wait for the next year's course for a more systematic treatment of the problem. They must be satisfied with studying one particular case, classical linear models. Given the historical baggage that they will encounter in the literature of any field, this really is necessary.

5.1.1 More assumptions or more data

The important point to get across in this chapter is that, in spite of appearances, we are still doing the same thing as in Chapter **??**: studying how the form of histograms changes under varying conditions. See Figure 5.1 in the text. However, the models make more assumptions because the histogram must here have the form of the normal curve. At the same time, this simplifies things, because only the mean will change, instead of each bar of the histogram (almost) independently.

Because the methods in Chapter **??** were so generally applicable, students may feel that they are sufficient. (In many cases, they are right.) They should be brought to consider why introducing a density function may be useful. From the results of describing histograms in Chapter **??**, this should be fairly obvious. Among other things, we find again smoothing and simplicity, as well as obtaining information on how the data might have been generated. An additional reason is that these models can be used even when sufficient observations are not available to construct frequency tables as in Chapter **??**.

5.1.2 Generalised linear models

This family of models was revolutionary in the 1970s. At present, it is a shackles on most statisticians who cannot see just how limited it is! Again, students need some familiarity with it in order to face the literature.

5.1.3 Location regression models

The existence of this family shows that generalised linear models need not be the only solution to the normal distribution's limitations. However, neither of these families even includes the so widely used Weibull distribution!

5.2 Linear regression

The way in which inferences are made in this chapter will depend on the instructor. I present both the classical tests and direct likelihood methods based on the AIC and like to give the students the choice of which they use. (It is not difficult to guess what is almost invariably chosen.) It is important to show that, although the ANOVA tables look more complex, the calculations are essentially the same, involving the same sums of squares.

5.2.1 One explanatory variable

Note that the AIC is here defined slightly differently than for categorical data. The deviance at the top of page 240 does not compare the model of interest to a saturated model (which would have a variance of zero), but instead to a maximal model of interest. Instead of adding $2p$ to this deviance, the latter is decomposed into its two parts, corresponding to the two models being compared. Twice the corresponding number of parameters in each model is then added to each part. One inconvenience is that values may be negative. However, differences in AIC, the only comparison of interest, will be identical to any other definition of the AIC.

For more advanced classes, it may be useful to mention how to make the AICs in this chapter comparable with those obtained using non-normal distributions. We need to define the AIC for the model of interest as

$$-2\log[\Pr(y_1,\ldots,y_n)]+2p$$

which contains the normalising constant and the unit of measurement Δ_i in its definition. Again, illustrating this idea is probably outside the scope of this introductory course, although students should understand after having studied the material in Chapters **??** and **??**.

The link between least squares calculations for linear regression and the methods for logistic and log linear regression in Chapter **??**, mentioned in passing in the text, should be brought out. The calculations for the estimates on pages 82, 97, and 238 are similar, the only essential difference being the weights, w_i, used. Here, they are identically one.

The essential thing with the babies' weights example is to link together Figures 5.1 and 5.2. Linear regression is not about fitting a straight line through data! It is about how the form of a distribution (now tightly constrained), represented by the histogram, changes with an explanatory variable.

The linear regression example on the effects of dieting is purposely chosen to illustrate a case where the null model of interest does not have a zero slope. No one would expect that weight after a diet was independent of weight before. This example also provides a clear opportunity to discuss the difference between one- and two-sided tests, if the instructor considers this to be important. However, linear regression is certainly not the best way to analyse these data; a paired means analysis would be more appropriate, as we shall see later in the chapter.

In explaining how a normally-shaped histogram is positioned along the fitted regression line for the mean, it is useful to calculate an interval about the line of say two standard deviations (as in Figure 5.2). This can be done using the tools of Section **??**. (Emphasise that it only works for that distribution, because the variance can be assumed to be constant.)

With a variance estimated as $\hat{\sigma}^2 = 4.190$, the estimated standard deviation is about 2. For people having a given prior weight of x_i, the weights after diet will have a mean of $4.187 + 0.910x_i$. About 95% of them will be distributed between $0.187 + 0.910x_i$ and $8.187 + 0.910x_i$, found by subtracting and adding two standard deviations. Thus, for example, people with prior weight of 64 kg will, on average, be 62.4 kg after diet, but the 95% range will be 58.4 to 66.4. This shows how the model is taking into account individual human variability, indicating that a substantial proportion gain weight although there is an average weight loss.

For both examples, most of this is based on theoretical assumptions that cannot be checked with only 24 or ten observations. Thus, in the second example, the students should be able to see that they cannot construct a histogram of weights after diet for each given prior weight. Discuss with them what would be required in order to check the model. They should see that they would need a sample with a considerable number of people for each prior weight of interest. With such data, they would probably discover that the distribution of posterior weights, for each given prior weight, was not normal but skewed (weights, unlike heights, usually are - why?), possibly close to one of those already studied in Chapter **??**. This will imply that the variance is not constant as well. In addition, one might also discover that the mean does not follow a straight line.

However, although none of these assumptions can be checked with only so few observations, useful information, if approximate, about the relationship between babies' weights and age or weight before and after diet may have been obtained. Remind the students that all models are approximations to reality in any case.

5.2.2 Multiple regression

Multiple regression (with an arbitrary distribution) is central to statistical modelling. Thus, this small section is important.

In the example, the idea of coding a binary explanatory variable as 0 and 1 is slipped in. The students should already be very familiar with constraints on qualitative explanatory variables in Section **??**. Now they learn how it can be implemented in practice with multiple regression. Emphasise that this also works with logistic and log linear models.

In this way, dummy variables are introduced to 'show' that ANOVA and regression are really the same thing. The former is just a special case of the latter.

5.3 Analysis of variance

5.3.1 One explanatory variable

ANOVA should be simple for the students, after all the work in Chapter **??**: (usually) no logarithms!

After blindly attacking the data on times taken to do homework in Chapter **??** using a normal distribution, the results of that chapter are used to decide to transform the data. This shows how log normal regression can also be done by these techniques. However, care should be taken in the presentation: otherwise, this can lead to some confusion, the danger being that students think that the logarithm of the response should be taken for any data set when applying this ANOVA.

The most common one-way analysis, difference of two means, is taken as a special case of ANOVA. In these ways, the students, as throughout the course, should be brought to see that these are not *ad hoc* recipes but a unified approach to modelling data.

5.3.2 Two explanatory variables

ANOVA is widely used in psychology, hence the example with two explanatory variables concerning learning scores. I do not know if Koerth pursuit rotor scores actually are constructed to be vaguely normal, and there are not sufficient data to check. Because the interaction can be taken to be zero, the sums of squares for residual and interaction can be combined to obtain a new residual sum of squares. This is equivalent to the standard likelihood comparison using the no interaction model as a basis of comparison for eliminating main effects. It is also a direct rationale for using the interaction sum of squares as residual when there is only one observation for each combination of explanatory variables.

Notice that the sums of squares for the main effects can be most easily obtained as sums of squares of the parameter estimates instead of recalculating the differences between means.

5.3.3 Matched pairs

Matched pairs is an important type of design for obtaining information efficiently in many fields. Continuation of the dieting example should allow the student to see that developing an appropriate model requires clear reasoning about what is going on. The presentation here should contrast with the usual view that matched pairs simply require an appropriate test.

5.3.4 Analysis of covariance

We have already encountered analysis of covariance in the multiple regression model for babies' weights above. Here, a rather old-fashioned presentation is given. However, the advantage is that it brings out the links between linear regression and ANOVA and serves as a very clear illustration of what interaction is all about. It also provides some of the basic steps in any model selection procedure. Because of the previous example, no new example has been provided (but there are exercises for the students).

5.4 Correlation

All students seem to have heard of correlation. By now, they should have realised that there are better ways to look at association and dependencies among variables in most contexts (especially for categorical variables). Correlation will, however, be used in Chapter **??**.

The presentation of the correlation coefficient is only the second place in the text where multivariate observations are discussed (the first was for log linear models). Correlation and simple linear regression can be contrasted by showing that the first must be represented as a simultaneous two-dimensional histogram and the second as a moving one-dimensional one.

The important point with correlation is that conclusions only be drawn in the wrong direction: no dependence means zero correlation but zero correlation does not mean lack of dependence. Give examples: the dependence may be non-linear. Show that the same problem occurs with simple linear regression, but is easier to get around, because, for example, a quadratic term can be added.

5.5 Sample size calculations

Here the general theory of likelihood-based sample size calculations, outlined in Chapter **??**, is applied to normal distribution to yield the same results as from a more classical power calculation.

5.6 Solutions to the exercises

Question (1)

The times in seconds for 30 children, classified by age, to push a hockey ball between a series of sticks during physical education were recorded (McPherson, 1990, p. 272) as shown in the following table:

Age	Time							
10	37	45	41	87	53	27	105	46
	27	35	38	54	19	36	30	
16	9	14	11	14	9	18	6	8
	30	8	10	12	16	23	14	

(a) Is there any indication of a difference in time between the two age groups?

(b) Do you think that the variability is the same in the two age groups?

(c) Why is a model for matched pairs not suitable for these data?

Answer

(a) The ANOVA table is

	SS	df	MSS	F
Age	7616.1	1	7616.1	26.9
Residual	7935.1	28	283.4	

The AIC with no difference in mean is 276.7 whereas it is 258.5 with different means and constant variance. Both show clear evidence of a difference in mean with 45.3 sec for age 10 and 13.5 sec for age 16.

(b) Yes, there is a difference in variability. The variances are respectively 491.16 and 37.85. The AIC with different variances is 240.6.

(c) Different children are involved in each group. There is no pairing among the values in the two age groups. Either set of times could be reordered without changing the results.

Question (2)

The table below gives the percentage of eligible voters casting ballots in the 1964 Vancouver civic election and the mean income (dollars) in 1961 in 24 districts of the city (Erickson and Nosanchuk, 1977, p. 206).

District	Income	Turn-out
East side		
Cedar Cottage	3974	40
Collingwood	4186	38
Fraserview	4173	42
Grandview	3864	42
Kingsway	3865	38
Little Mountain	4383	43
Mt. Pleasant	3422	30
New Brighton	4003	39
Newport	4594	41
Riley Park	3865	38
Strathcona	2751	24
Sunset	4299	40
Woodland	3315	26
West side		
Arbutus	6267	55
Burrard	3589	27
Dunbar	5701	58
Fairview	3786	30
Kerrisdale	7066	59
Kitsilano North	3785	34
Kitsilano South	4558	41
Marpole	4640	41
Pt. Grey	5908	48
Shaughnessy	8477	52
West End	4233	33

The districts of Vancouver are distinct in the eyes of long-term residents of the city, and almost anyone who lives there knows them by name, although they have no administrative status.

(a) Study the relationship between the two variables, income and turn-out. Graphics will be useful.

(b) What reasons can you find to explain your results?

(c) Suggest a better model if the number of eligible voters in each district were available.

(d) The unit of observation is the district, each having a distinct geographical location. Is it reasonable to assume that such observations are independent?

(e) Is this a sample, and is there a well-defined population?

(f) Does it make sense to draw inferences about true parameter values in a model for data such as these?

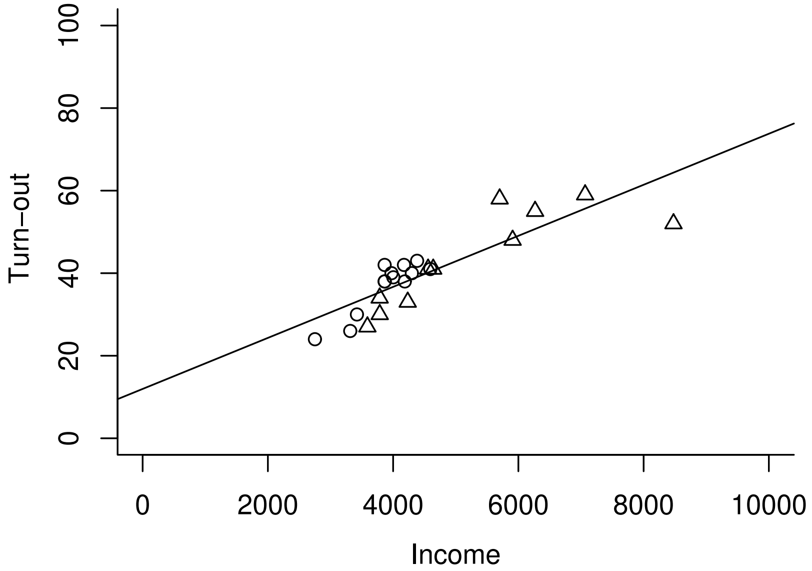

Figure 5.1: Scattergram for voting behaviour in Vancouver (circles for the east side, triangles for the west side), with the fitted linear regression. (Erickson and Nosanchuk, 1977)

Answer

(a) The data are plotted in Figure **??**. They appear to follow a fairly straight line, with the exception, perhaps, of the richest district, Shaughnessy, which has a low turn-out. We may fit a linear regression, giving $\hat{\beta}_0 = 11.95$ and $\hat{\beta}_1 = 0.00618$. This is also plotted on the graph. Turn-out increases, on average, by 0.6% for every $100 increase in income.

The deviance, from Equation (5.2), comparing models without and with income, is 30.18 with one degree of freedom. The corresponding AICs are 108.82 and 80.64, respectively, showing that the model with dependence of mean turn-out on income is superior to that without. The Student t test gives $t_{22} = 7.44$ which yields a P-value, on a two-sided test, less than 0.001, so that the hypothesis of no effect is clearly rejected.

(b) This type of data is called ecological because it contains no information about individuals but only about groups. We would require knowledge of political science and of voting habits in local Vancouver elections, such as information about the parties and candidates running and the issues at stake, in order to venture conclusions about why there is higher turn-out in higher income districts.

(c) If we had the number of eligible voters in each district, we could try fitting a logistic regression to these data.

(d) Candidates, and perhaps issues, will be different in the various districts. People may have chosen to live in a district because they find similar people there. Districts closer together may often be more similar than those further apart.

(e) All districts in Vancouver are represented, so that this is not a sample for the 1964 election. It might be considered to be a sample from the series of civic elections in different years in Vancouver, but it is certainly not a *random* sample.

(f) With some imagination, we might argue that this analysis tells us something about other elections in Vancouver, before or after, but candidates and issues will surely change, making the conclusions doubtful. The regression model describes these data, with little justification for extrapolation. Thus, the deviance and AIC show which model fits best, but the test has little meaning.

Question (3)

The table below presents the murder rates in 24 randomly chosen cities in the U.S.A., classified by type and location (Blalock, 1972, p. 335).

	City type					
Region	Industrial		Trade		Government	
NE	4.3	5.9	5.1	3.6	3.1	3.8
	2.8	7.7	1.8	3.3	1.6	1.9
SE	12.3	9.1	6.2	4.1	6.2	11.4
	16.3	10.2	9.5	11.2	7.1	12.5

(a) Study the effects of each of these two nominal variables separately on the observed murder rates.

(b) Now construct a model to describe the simultaneous effects of the two explanatory variables.

(c) Discuss the advantages of this second approach over the first.

(d) Rates refer to the occurrence of events, here murders. If we had the numbers of such events in each city, what other information would we require in order to calculate the rates?

(e) What distribution might be more appropriate than the normal, if such information were available?

Answer

(a) For type of city, the estimated mean murder rates are 8.58, for industrial cities, 5.60 for trade, and 5.95 for government. There appears to be little difference between the latter two types. The deviance is 2.88 with AIC 68.99 as compared to 67.88 for the model without the variable, type of city, indicating no evidence of difference among types. The F test gives $F_{2,21} = 1.34$ with a P-value greater than 0.20, agreeing with the AIC and the deviance test conclusion.

For regions, the estimated rates are 3.74 for NE and 9.68 for SE. The deviance is 20.00 with AIC 49.88, (as compared to the same AIC as above, 67.88, without region) indicating that there is a difference between regions. The F test is $F_{1,22} = 28.63$ giving (similarly to the deviance test) a P-value less than 0.001, confirming this conclusion.

(b) A model with both explanatory variables will also contain the interaction parameters. The parameter estimates, using the mean constraint, are $\hat{\mu} = 6.708$, $\hat{\alpha}_1 = 1.867$, $\hat{\alpha}_2 = -1.108$, $\hat{\alpha}_3 = -0.758$ (for type), $\hat{\beta}_1 = -2.967$ (for region), $\hat{\gamma}_{11} = -0.433$, $\hat{\gamma}_{21} = 0.817$, and $\hat{\gamma}_{31} = -0.383$.

The analysis of variance table is

Effect	SS	MSS	d.f.	F
Total	373.54	16.24	23	
City type	42.30	21.15	2	3.53
Region	211.23	211.23	1	35.21
Interaction	8.01	4.01	2	0.64
Residual	112.00	6.22	18	
Residual+Interaction	120.01	6.00	20	

This clearly indicates that the interaction can be eliminated. Therefore, the interaction and the residual lines in the ANOVA table can be added, yielding the last line in the table; the resulting mean sum of squares obtained from this is used as denominator of the F-statistics for the main effects. The F test for difference in type of city has a P-value between 0.05 and 0.02, and that for region is very significant. The model with both effects but without the interaction has an AIC of 46.63, whereas the full model has 48.97. When compared to the AICs previously given, this indicates that the model with both region and type of city is preferable.

(c) The first approach only gives us the average relation for type of city, ignoring region, and of region, ignoring type of city. It does not allow us to determine if the type of city has a different relationship to the murder rate in the two regions.

(d) If we had the number of such events in each city, we would also require the population of each city in order to calculate the rates.

(e) If the numbers of events were available, a distribution for counts, such as the Poisson or negative binomial would be more appropriate. However, this would have to be corrected by allowing for the population sizes, by using an offset in the log linear model. See the answer to Exercise **??**.

Question (4)

The data in Exercise (**??**.**??**) above are classified by the two main regions of Vancouver, the East and West sides. Construct models to compare

(a) the mean income for the two regions;

(b) the percentage turn-out for the regions.

Answer

(a) For mean income, the one-way analysis of variance has means of \$3900 and \$5274, respectively, for the east and west sides. The deviance, comparing without and with difference, is 8.29, the AICs being 344.92 and 336.62 respectively. This indicates a difference in income between the two sides. The F test gives $F_{1,22} = 9.08$ for a P-value between 0.01 and 0.001, indicating rejection of the hypothesis of no difference. (This can also be calculated as a Student t test, $t_{22} = 3.01$, with the same conclusions, of course.) A Chi-square test with one degree of freedom based on the deviance also leads to the same conclusion.

(b) For voting turn-out, the difference between the two sides can be seen in Figure **??**. The one-way analysis of variance has means of 37.0% and 43.5%, respectively, for the east and west sides. The deviance, comparing without and with difference, is 3.09, the AICs being 108.82 and 107.74 respectively. This provides a small indication of a difference in turn-out between the two sides. The F test gives $F_{1,22} = 3.02$ for a P-value between 0.10 and 0.05, indicating that the hypothesis of no difference is not rejected. The test based on the deviance yields the same result. Here, the AIC and the tests are in some disagreement. With one degree of freedom, tests point to a simpler model than the AIC. (The reverse will be true for degrees of freedom greater than seven.)

Question (5)

Construct a model to compare the relationship between voter turn-out and mean income for the two regions of Vancouver in Exercise (**??**.**??**) above.

Answer

We can use a model for analysis of covariance. We have already fitted three of the simpler models above. Recall that the AICs were 108.82 for the null model, 80.64 for that with only income, and 107.74 for that with only the region of the city. That for parallel regression lines in the two regions has 81.15, whereas that with two distinct regressions has 79.95. Thus, the (very slightly) preferable model is this last one, although it is not much better than the one with only income according to the AIC.

For the east side, the parameter estimates are $\hat{\beta}_0 = -7.31$ and $\hat{\beta}_1 = 0.0114$, whereas they are $\hat{\beta}_0 = 10.52$ and $\hat{\beta}_1 = 0.00625$ for the west side. The model is plotted in Figure **??**. We see how the lower turn-out for high income districts in the west side has pulled down that regression line.

Question (6)

The following table gives the salaries (dollars) of board chairpersons of community organisations in the U.S.A., classified by type of organisation and size of community (Blalock, 1972, p. 358).

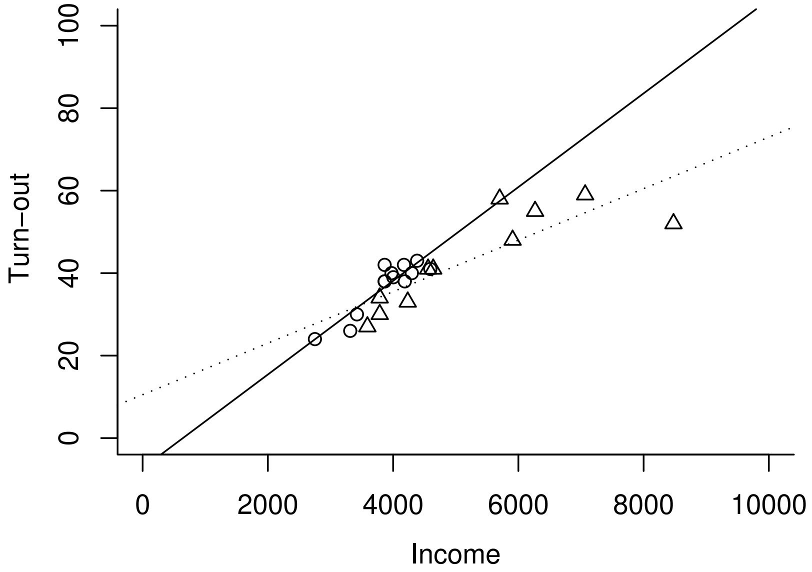

Figure 5.2: Scattergram for voting behaviour in Vancouver (circles for the east side, triangles for the west side), with the separate fitted linear regression lines for the east (solid) and west (dotted) sides. (from Erickson and Nosanchuk, 1977)

Size of	Organisation type		
community	Religious	Social welfare	Civic
Large	13000	15000	20800
	11500	10600	18100
	17300	12300	18100
	19100	11400	22300
	16700	10800	16500
Small	15000	9300	14400
	12300	10400	10800
	13900	12900	9700
	14300	11000	12300
	11700	9100	13100

Five organisations of each type were randomly selected for both large and small communities. No further information is given about the definitions of the variables. Study the relationships between these two nominal variables and the salaries by fitting an appropriate model.

Answer

We shall fit a two-way analysis of variance model. The AICs are 464.19 for the full model, 471.37 for that without interaction, 482.55 for that with only size of community, 483.81 with only organisation type, and 490.06 with neither. This indicates that the full model, with interaction is necessary. For comparison, the ANOVA table is

Effect	SS	MSS	d.f.	F
Total	348707000	12024379	29	
Size	94696330	94696330	1	21.54
Type	100886000	50443000	2	11.47
Interaction	47620670	23810330	2	5.42
Residual	105504000	4396000	24	

The F test for the interaction being zero is significant so that this confirms that the full model is required.

The parameter estimates, using the mean constraint, are $\hat{\mu} = 13790.0$, $\hat{\alpha}_1 = 690.0$, $\hat{\alpha}_2 = -2510.0$, $\hat{\alpha}_3 = 1820.0$ (for type), $\hat{\beta}_1 = 1776.7$ (for size), $\hat{\gamma}_{11} = -736.7$, $\hat{\gamma}_{21} = -1036.7$, and $\hat{\gamma}_{31} = 1773.4$. On average, salaries are higher in large communities and in religious organisations. However, they are relatively much higher for civic organisations in large than in small communities.

Question (7)

The following table gives estimations of an index of the cost of living in five areas of Bengal, India, in 1945 by five investigators (Yule and Kendall, p. 529):

	Area				
Investigator	A	B	C	D	E
1	270	263	264	263	260
2	280	265	274	274	279
3	275	284	278	271	296
4	271	269	272	297	274
5	279	267	269	263	284

(a) Is there a difference in cost of living among the areas?

(b) Is there any evidence that the investigators differ in their evaluations of the areas?

(c) Why can you not determine if each investigator used the same criterion in all areas?

Answer

(a) The ANOVA table is

	SS	df	MSS	F
Investigator	775.4	4	193.8	2.6
Area	239.0	4	59.7	0.8
Residual	1175.4	16	73.5	

The AIC with no difference in mean is 186.8, whereas it is 183.8 with difference in investigator, 191.9 with difference in area, and 187.2 with difference in both. Thus, neither approach indicates a difference among areas. The means are 264.0, 274.4, 280.8, 276.6, and 272.4.

(b) On the other hand, the ANOVA table indicates no differences at the 5% level, whereas the AICs indicate a difference in investigators.

(c) The interaction between investigator and area cannot be estimated because there is only one observation for each investigator/area combination.

Question (8)

In a study of 24 fifth-grade children at the School of Behavioural Sciences in Macquarie University, Australia, the time taken to solve four block design problems and the value for the embedded figures test (EFT), a measure of difficulty in abstracting logical structure of a problem from its context, were recorded. The children were classified by the type of problems presented first, those solved by row (group 1) or by formation strategy (group 2) as shown in the following table (Aitkin *et al.*, 1989, p. 344):

Group	Time	EFT	Time	EFT	Time	EFT	Time	EFT
1	317	59	464	33	525	49	298	69
	491	65	196	26	268	29	372	62
	370	31	739	139	430	74	410	31
2	342	48	222	23	219	9	513	128
	295	44	285	49	408	87	543	43
	298	55	494	58	317	113	407	7

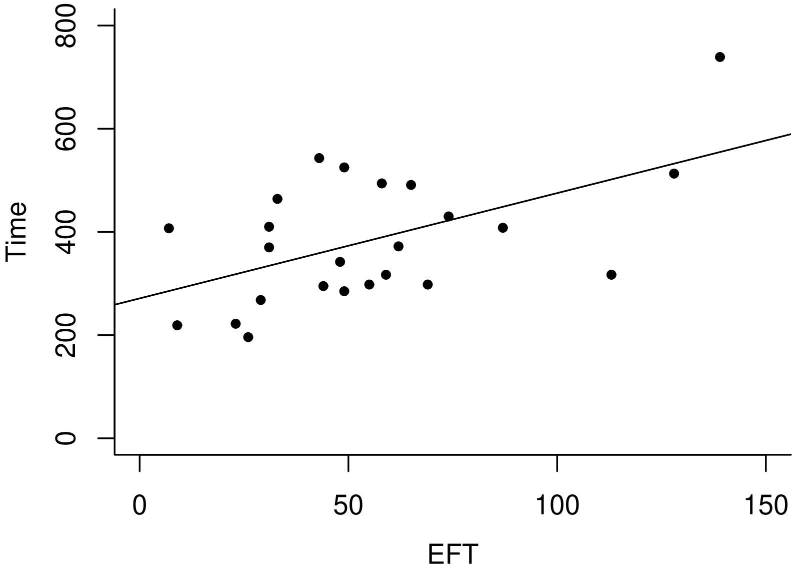

Figure 5.3: Scattergram for the dependence of time on EFT, with the fitted linear regression line.

(a) Is there any relationship between the time taken on the block design problems and the results of the embedded figures test?

(b) Do the results for either of these measures differ with the order of presentation?

(c) Develop a complete model for these data and explain your conclusions.

Answer

(a) The AIC for the null model is 303.4 as compared to 296.8 when time depends on EFT, indicating a relationship. The equation is

$$\mu_i = 271.13 + 2.04x_i$$

where x_i is the EFT score. This is plotted in Figure **??**.

(b) No. The AICs for time, without and with an order effect are 303.4 and 304.6. Those for EFT are respectively 240.3 and 242.3.

(c) The above model for dependence of time on EFT is sufficient. Adding group and the interaction between group and EFT does not improve the model. However, using a log transformation (log normal distribution) does improve the model somewhat: AIC 294.7, not surprising given that these are durations. (The gamma distribution also provides about the same fit.)

Question (9)

(a) Would any of the explanatory variables in any of the exercises above be susceptible to use as causal effects?

(b) Do the above data tables provide such information?

(c) How would you go about collecting information on causality in such contexts?

Answer

(a) The only explanatory variable in the above exercises that was controlled was the order of the problem presented in Exercise (**??**.**??**).

Consider, however, the example of babies' weights. There, it might be possible to control is gestation age, for example by having a mother avoid strenuous exercise to prevent premature birth. Thus, this might possibly be considered as a causal variable.

(b) We have no information on how the data on birth weights were collected but it is certain that a controlled experiment was not performed, setting the weights at random! Thus, one must be extremely cautious in drawing any conclusions about increasing birth weight by trying to increasing gestation age. One aspect favourable to the possibility of such a conclusion is that the relationship is the same for both sexes.

(c) An experiment might be designed whereby some mothers-to-be in danger of giving premature birth were randomly chosen to follow a treatment believed to increase gestation age and another group not. The birth weights of the babies would then be recorded and the results analysed as we have done. In such an experiment, the conclusions would be much stronger than those we have drawn, although we should not expect the actual model estimates necessarily to be similar.

Question (10)

Many of the data sets in the examples and exercises of this chapter may seem rather contrived. This illustrates the very real difficulty in finding data which might plausibly be described by models based on the normal distribution. Which of these tables do you think contain purely invented data?

Answer

The tables on murder rates and salaries appear to be pure inventions. In any case, it seems strange that the same numbers of observations appear in each category. Little information is provided about the source of the data except to say that the cities and the organisations were randomly chosen.

Question (11)

(a) Calculate the sample size required to detect a difference in means of 10 for a suitable value of the deviance.

(b) Plot this as a function of the variance.

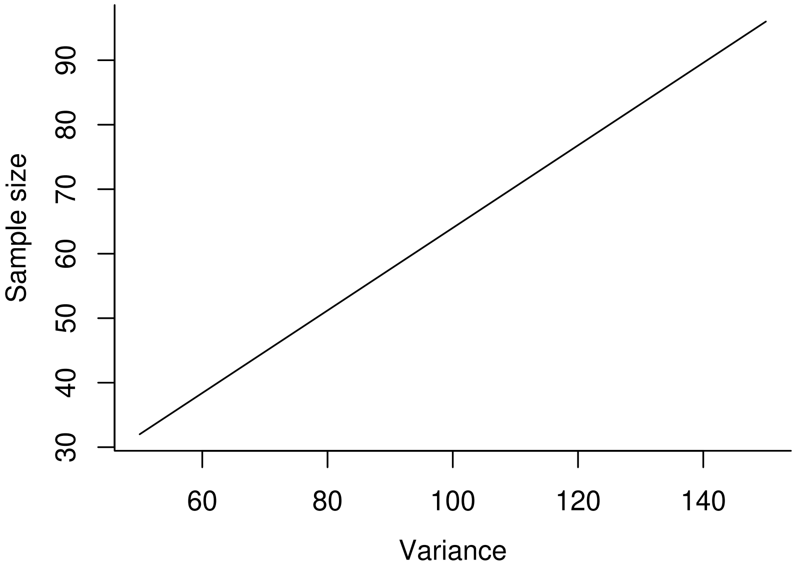

Figure 5.4: The relationship between sample size and variance for a difference of means of 10 from normal distributions.

Answer

(a) Suppose first that the variance is 100. We take a deviance of four. Then the sample size required is 64.

(b) The sample sizes for various values of the variance are plotted in Figure **??**. We see how they increase linearly with the variance.

Chapter 6

Dependent responses

The goal in this short chapter is to try to show students some of the wide areas for more complex applications of statistical models. These should be appealing in almost any field, from economics (time series) to medicine (survival curves). Special effort has been made to choose models that require no new material. Thus, this chapter provides applications of the methods of the preceding chapters. Many of the concepts rather incidentally introduced in previous chapters reappear here: assumptions (such as the Markov) made to simplify models, the intensity function, unmeasured heterogeneity (as in overdispersion), and so on.

6.1 Repeated measurements

It will be useful to point out to the students that the models used in this chapter involve multivariate responses. A series of observations on an individual will be dependent, requiring such models. This contrasts with virtually all of the models in previous chapters, where independent observations were assumed.

6.2 Time series

6.2.1 Markov chains

Point process data can be looked at in a number of ways. We saw this in Chapter **??** when studying the Poisson and exponential distributions. The graph in Figure 6.1 of the text is a useful representation of this. The students should discuss ways of applying the methods of that chapter, such as creating a frequency table of the number of different lengths of times between accidents or looking at the number of accidents each week or month.

I have chosen to look at another, perhaps less obvious (to the students) way of handling such data: Markov chains. This answers an important question about dependence among events at successive points in time. Once the tabulation of the table is accomplished (tedious without a computer), the modelling is very simple, but effective. The results in the example, that accidents depend on what happened two days, but not one day, before will surely bring discussion from the students.

A Markov chain is auto-regression for a categorical response. This is introduced first, before the classical auto-regression because it is simpler and easier to understand, given what has gone on previously in the course.

6.2.2 Autoregression

One simple example of classical auto-regression, with a normal distribution, here serves to introduce a number of important concepts: stationarity, drift, trend, random walk. This is a first application of correlation, appropriate because time series are multivariate observations.

6.3 Clustering

Random effects models provide a second example of an application of correlation. This time the multivariate observations are clustered rather than sequential. The explanation of the model is heuristic, involving a lot of hand-waving. An interesting point to discuss is that heterogeneity can only be detected because several observations are available on each individual. Heterogeneity of responses across individuals is just the other side of the coin to homogeneity of responses on the same individual. Total variability is made up of the sum of the two.

As we saw in Section **??**, in certain disciplines, clustering is built into many study designs, for example, when the unit of random sampling is a group, such as a classroom, family, or village, but each member of the group is questioned. Another important application of clustering is to meta-analysis. This involves the secondary analysis of a series of similar studies in order to combine the information from them. Responses within a given study may be expected to be more similar than those across different studies, so that each study is the ‘individual’ forming a cluster.

In sophisticated classes, it will be interesting to discuss what happens if repeated responses on individuals are more heterogeneous than across individuals. From the formula, it is obvious that the intraclass correlation could be negative. This could happen if there is some sort of repulsion among events on an individual so that they are forced to be quite different.

6.4 Life tables

6.4.1 One possible event

The construction of a life table is essentially an exercise in reorganising data in a sensible way. It has the additional complication that there are really two variables being measured, the time and the censor indicator. A simple application of a logistic model allows us to fit the famous Cox proportional hazards model.

6.4.2 Repeated events

An important theoretical point, that usually has little practical relevance (as in the example), is the difference between the case when only one event at a time is possible, and when several may occur. Individuals can only die once, so that survival falls into the former case. Students can discuss other situations, including some of the previous examples in the text (accidents, divorces, consumer purchases).

6.5 Solutions to the exercises

Question (1)

The table below gives a series indicating if patients were arriving (indicated by 1) at the intensive care unit of a hospital in the Oxford, England, Regional Hospital Board each day from 4 February, 1963 to 18 March, 1964 (Lindsey, 1992, p. 26, from Cox and Lewis, 1966, pp. 254–255; read across rows).

00010	00100	10000	10101	10001	00110	10001	01000
00111	00101	01000	10100	10001	00111	00011	00000
01000	01100	00101	10001	01101	01110	11110	01010
10101	00001	01100	10100	11011	11011	01000	00111
01100	00001	10110	01010	01110	00100	01010	00001
01001	00000	01010	01011	01101	01101	00101	10011
00111	00101	00011	00000	11011	00100	01110	01111
11011	00111	11001	11011	01111	10101	11011	11111
00111	11100	10010	11011	10011	10110	10111	00110
00111	00001	11000	11000	01111	00111	10001	01010
00110	00000	1					

(a) Plot the cumulative number of events against time and interpret the resulting graph.

(b) Fit a model to determine if there is any relationship between patients arriving on successive days.

(c) Group the data by month and plot them to see if there is evidence of stationarity.

(d) What can be concluded about any systematic change over time? Consider both steady changes and seasonal effects.

Answer

(a) The cumulated number of patients is plotted in Figure **??**. Because the graph shows a relatively straight line, the proportion of days with patients arriving is constant over time. This proportion is indicated by the slope. Note that this is not a typical rate (number of patients per day) because we do not know how many patients arrived on a given day (unless it is zero).

(b) We can construct a two-way table showing a day with patients arriving, or not, is followed by a day with patients arriving, or not:

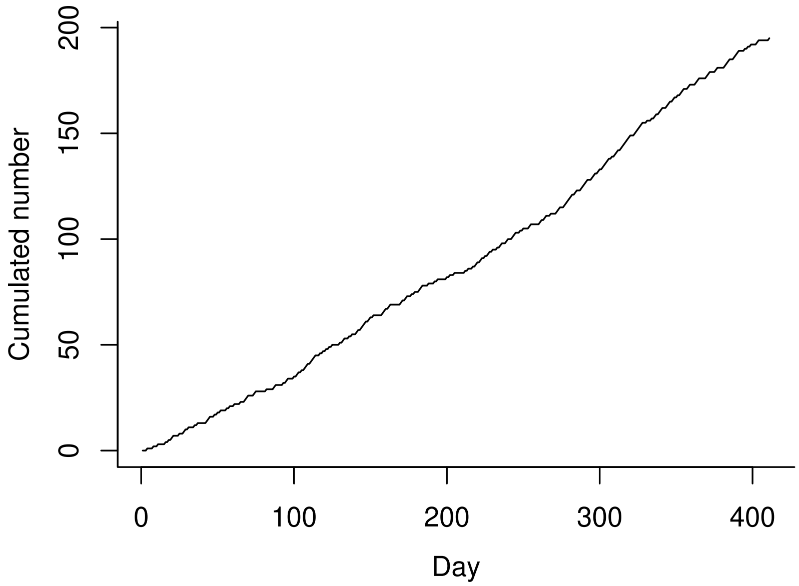

Figure 6.1: Cumulative number of patients were arriving at the intensive care unit of a hospital in the Oxford, England, Regional Hospital Board each day from 4 February, 1963 to 18 March, 1964. (Lindsey, 1992, p. 26)

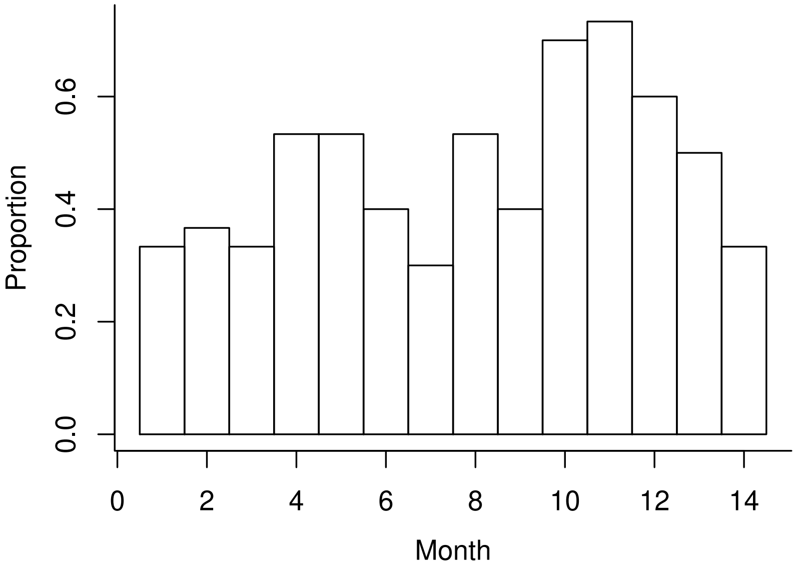

Figure 6.2: Proportion of days in each month when a patient arrived at the intensive care unit of a hospital in the Oxford, England, Regional Hospital Board each day from 4 February, 1963 to 18 March, 1964. (Lindsey, 1992, p. 26)

	Day t	
Day $t-1$	No	Yes
No	112	104
Yes	103	91

When we fit a simple logistic model for independence to these data, we obtain a deviance of 0.06, with AIC 2.06, as compared to 4 for the saturated model. Thus, there is no evidence for dependence between the probabilities of patients arriving on successive days. If we had the hypothesis of independence before obtaining the data and applied a significance test based on the deviance, then again there would not have been any real evidence of dependence, the P-value being greater than 0.2.

(c) To facilitate calculations, we cut the series into sequences of 30 days, as approximate months. We then calculate the proportion of days in which patients arrived and plot them as a histogram, shown in Figure **??**. More patients appear to arrive in the winter months of November, December, and January. This may indicate both a seasonal effect and/or a longer trend to overall increase, because February 1963 had fewer patients than February 1964 (although March of the two years is similar).

(d) We can study this more formally by fitting a logistic regression to the 14 months. The model of independence (constant probability for all months) has a deviance of 30.95 and AIC 32.95, whereas the regression on month has deviance of 24.28 (28.28) and the full model has an AIC of 28. This indicates that the proportion of patients entering per month is not random but is varying over the period. There is a large linear trend, but the full model with a different probability each month is slightly superior. (A closer look at Figure **??** reveals that the slope may be increasing in the later part of the period.) Of course, this is a rather short series, given that there is evidence of seasonal effects! It would be useful to know more about the type of problem each patient has. If these include respiratory problems, for example, a seasonal model would make sense.

Question (2)

Exercise (**??**.**??**) gave the traffic violations each year among male subjects in a driver education study. Develop a Markov chain model to describe these data.

Answer

In the Markov chain model, the result each year only depends on that the previous year. In Exercise (**??**.**??**), we saw that this model does not fit well because there are more long term dependencies. The AIC is 74.7 as compared to 27.3 for the best model found in that exercise.

Question (3)

Beveridge (1936) gives the average rates paid to agricultural labourers for threshing and winnowing one rased quarter each of wheat, barley, and oats in each decade from 1250 to 1459. These are payments for performing the manual labour of a given task, not daily wages. He obtained them from the rolls of eight Winchester Bishopric Manors (Downton, Ecchinswel, Overton, Meon, Witney, Wargrave, Wycombe, Farnham) in the south of England. As well, he gives the average daily wages of carpenters and masons in Taunton manor, and the average price of wheat for all England, as shown below.

Agriculture	Carpenter	Mason	Wheat price
3.30	3.01	2.91	4.95
3.37	3.08	2.95	4.52
3.45	3.00	3.23	6.23
3.62	3.04	3.11	5.00
3.57	3.05	3.30	6.39
3.85	3.14	2.93	5.68
4.05	3.12	3.13	7.91
4.62	3.03	3.27	6.79
4.92	2.91	3.10	5.17
5.03	2.94	2.89	4.79
5.18	3.47	3.80	6.96
6.10	3.96	4.13	7.98
7.00	4.02	4.04	6.67
7.22	3.98	4.00	5.17
7.23	4.01	4.00	5.45
7.31	4.06	4.29	6.39
7.35	4.08	4.30	5.84
7.34	4.11	4.31	5.54
7.30	4.51	4.75	7.34
7.33	5.13	5.15	4.86
7.25	4.27	5.26	6.01

(a) Fit an autoregression model to each series.

(b) Compare the results with those for time trend models.

(c) Does the price of wheat display any relationship to the rates paid to agricultural labourers?

(d) Fit a multiple regression model for wheat prices containing an autoregression, a time trend, and a dependence on agricultural rates.

i Are all these variables necessary in the model?

ii Interpret the results.

iii Plot profile likelihood functions for all important parameters.

Answer

(a) The four series are plotted in Figure **??**. They do not appear to show much in common.

The fitted autoregression equations are

$$\begin{aligned}
\mu_t &= 0.39 + 0.96 y_{t-1} \\
\mu_t &= 0.46 + 0.89 y_{t-1} \\
\mu_t &= 0.09 + 1.01 y_{t-1} \\
\mu_t &= 5.65 + 0.06 y_{t-1}
\end{aligned}$$

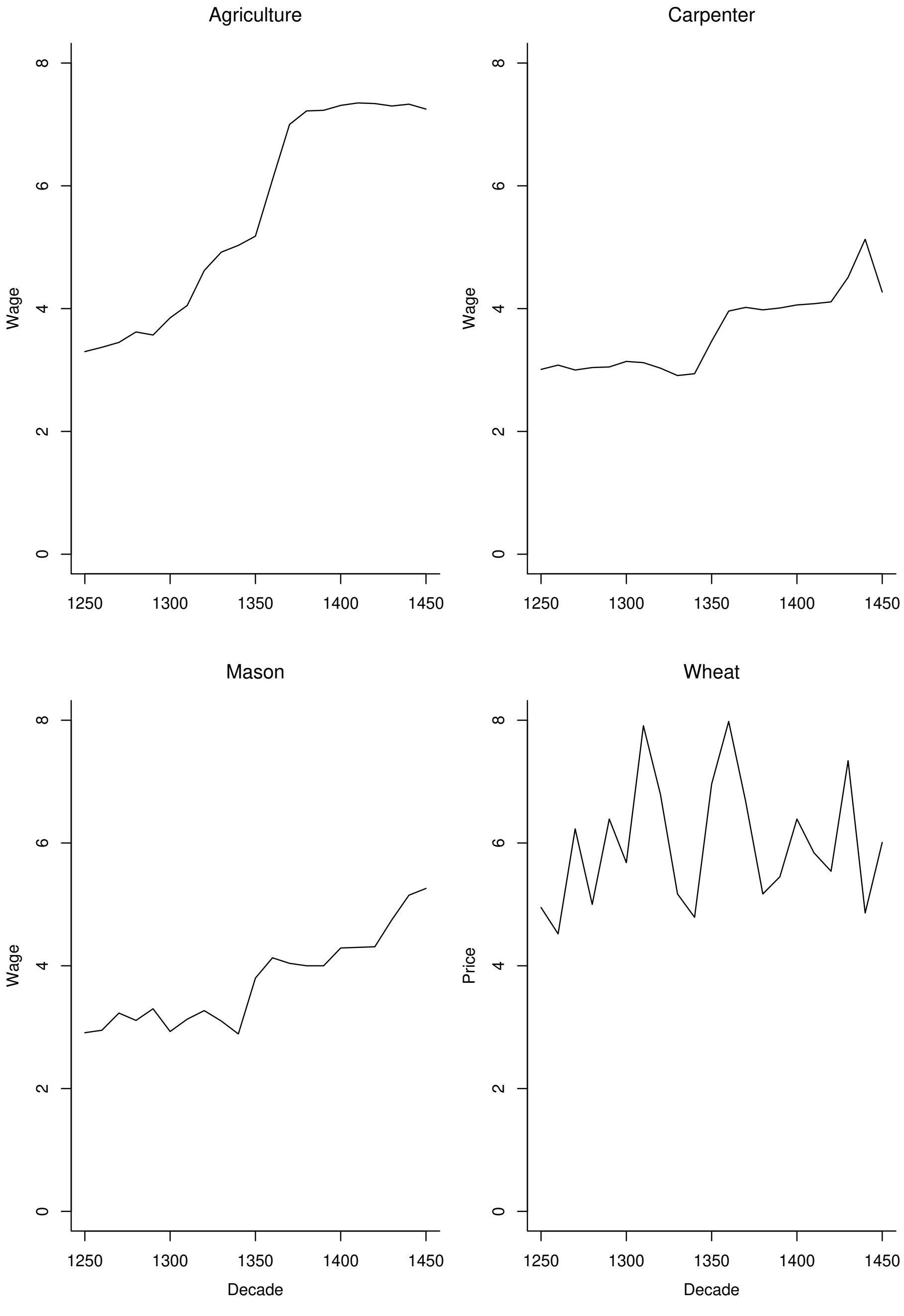

Figure 6.3: Average rates paid to agricultural labourers, carpenters, and masons, and the average price of wheat, each decade from 1250 to 1459. (Beveridge, 1936)

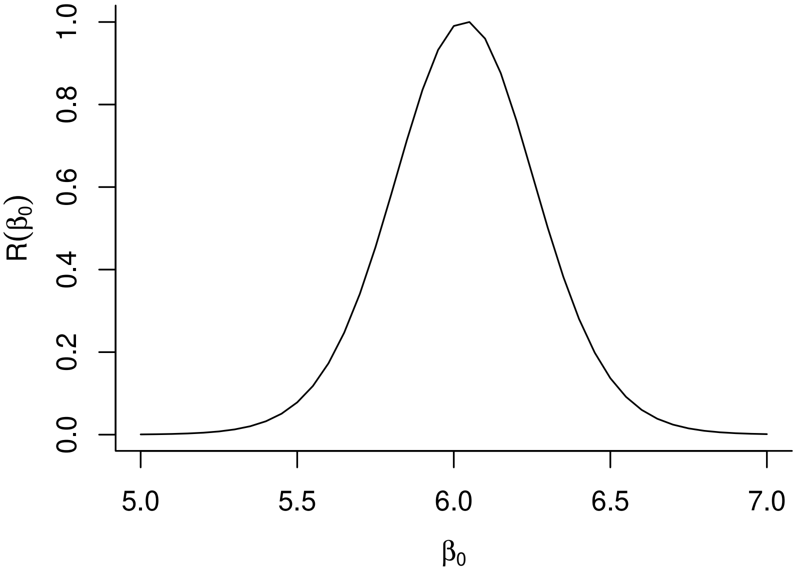

Figure 6.4: Normed likelihood function for the average price of wheat over the period from 1250 to 1459.

respectively, for agricultural labourers, carpenters, masons, and wheat prices, with AICs 10.7, 13.3, 11.2, and 62.3, as compared to the null models with 60.4, 42.0, 47.7, and 60.4.

(b) The corresponding regression equations for a time trend are respectively

$$\begin{aligned}\mu_t &= 2.90+0.26t\\ \mu_t &= 2.63+0.10t\\ \mu_t &= 2.59+0.12t\\ \mu_t &= 5.92+0.01t\end{aligned}$$

with AICs 32.8, 11.5, 12.4, and 62.3. Thus, the autoregression fits better for agricultural labourers and masons, the time trend for carpenters, and neither fits well for the wheat prices.

(c) No, the price of wheat does not seem to depend on agricultural wages either the same year or the year before.

(d) None of the variables are necessary in the model. None of the available information seems to be able to explain the erratic changes in the price of wheat. The normed likelihood function for the mean price of wheat is plotted in Figure **??**. Plausible values lie in the interval (5.6, 6.4), with maximum likelihood estimate 6.03. On the other hand, the standard deviation of wheat prices in this model is 0.99 so that, for

Table 6.1: Annual percentage increase in average wages of white collar workers in the U.S.A., 1962–1979. (Lindsey, 1992, p. 122, from Nichols, 1983)

1961–62	2.8	1970–71	6.2
1962–63	2.7	1971–72	6.3
1963–64	2.7	1972–73	5.5
1964–65	2.2	1973–74	6.2
1965–66	2.9	1974–75	9.1
1966–67	4.5	1975–76	7.6
1967–68	5.1	1976–77	6.9
1968–69	5.5	1977–78	7.5
1969–70	6.2	1978–79	7.2

example, 95% of wheat prices over this period should lie in the range (4.05, 8.01). By now, the students should understand clearly the difference between these two intervals!

Question (4)

Table **??** gave the evolution of wage increases for low grade jobs; the table below gives a similar series of annual percentage increases in average wages of white collar workers in high grade jobs in the U.S.A., 1962–1979 (Nichols, 1983).

1962	3.5	1971	6.2
1963	3.7	1972	5.6
1964	3.5	1973	5.7
1965	4.2	1974	6.2
1966	4.2	1975	8.8
1967	4.1	1976	6.5
1968	4.7	1977	7.7
1969	5.9	1978	8.8
1970	6.4	1979	8.0

(a) Plot and compare the two series.

(b) Find a reasonable model for this series.

(c) Compare it to the results given for the other series.

(d) Redo the analyses for the two tables using a log normal distribution.

Answer

(a) The two series are plotted in Figure **??**. They follow each other fairly closely, although the high grade workers had larger increases at the beginning of the period.

(b) We shall try the same models as for the low grade workers in the text. The autoregression with drift has parameter estimates, $\hat{\beta}_0 = 1.325$, $\hat{\beta}_1 = 0.812$, and $\hat{\sigma}^2 = 0.929$. The AIC is 2.74. That without drift has parameter estimates, $\hat{\beta}_1 = 1.028$ and $\hat{\sigma}^2 =$

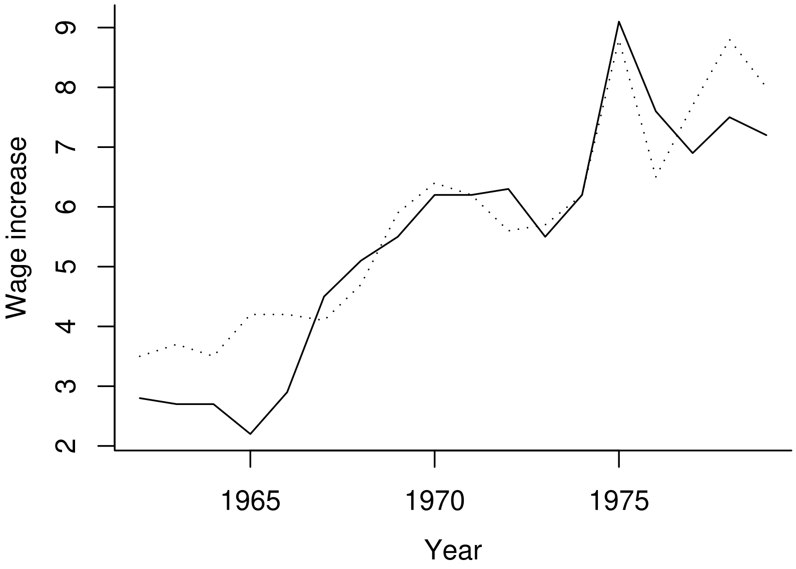

Figure 6.5: Annual percentage increase in average wages of low (solid) and high (dotted) grade white collar workers in Britain, 1962–1979. (from Nichols, 1983)

Table 6.2: Judgement of three different levels of albedos by four observers at illumination level 2. (McNemar, 1954, p. 321)

	Level of albedos		
Observer	0.07	0.14	0.26
1	14	24	65
2	27	36	47
3	18	24	62
4	24	59	84

1.068. The AIC is 3.12. The independence model has parameter estimates, $\hat{\beta}_0 = 5.894$ and $\hat{\sigma}^2 = 2.732$. The AIC is 19.09. Finally, the model with linear trend has parameter estimates, $\hat{\beta}_0 = 2.803$, $\hat{\beta}_1 = 0.309$, and $\hat{\sigma}^2 = 0.440$. The AIC is -9.97.

(c) These are very similar to the results obtained for the low grade workers, so that the conclusions will be the same.

(d) When we take the logarithm of the percentages, a common procedure for response variables in economics, we have to add $2\sum\log(y_i)$ to the AIC formula for the normal distribution to make them comparable. (This comes from the extra y_i in the denominator of the distribution; see page 117 of the text. Another way of looking at it is that this y_i is related to the unit of measurement, which after the log transform, becomes Δ_i/y_i.)

We obtain the following results. For low grade workers, the auto-regression with drift has parameter estimates, $\hat{\beta}_0 = 0.244$, $\hat{\beta}_1 = 0.881$, and $\hat{\sigma}^2 = 0.0284$. The AIC is 1.23. That without drift has parameter estimates, $\hat{\beta}_1 = 1.025$ and $\hat{\sigma}^2 = 0.0324$. The AIC is 1.47. The independence model has parameter estimates, $\hat{\beta}_0 = 1.638$ and $\hat{\sigma}^2 = 0.169$. The AIC is 29.54. Finally, the model with linear trend has parameter estimates, $\hat{\beta}_0 = 0.901$, $\hat{\beta}_1 = 0.0738$, and $\hat{\sigma}^2 = 0.0384$. The AIC is 6.36.

For high grade workers, the auto-regression with drift has parameter estimates, $\hat{\beta}_0 = 0.302$, $\hat{\beta}_1 = 0.850$, and $\hat{\sigma}^2 = 0.0196$. The AIC is -1.42. That without drift has parameter estimates, $\hat{\beta}_1 = 1.024$ and $\hat{\sigma}^2 = 0.0324$. The AIC is -1.25. The independence model has parameter estimates, $\hat{\beta}_0 = 1.734$ and $\hat{\sigma}^2 = 0.815$. The AIC is 20.83. Finally, the model with linear trend has parameter estimates, $\hat{\beta}_0 = 1.192$, $\hat{\beta}_1 = 0.0542$, and $\hat{\sigma}^2 = 0.0109$. The AIC is -11.39.

Here, the results are not as similar for the two grades. For the low grade, the auto-regression is superior to the trend model, giving a comparable fit to that without logarithms. However, the trend model without logarithms is the best of those tried. On the other hand, for the high grade, the trend gives a better model here and it is also superior to any model without logarithms. In summary, the trend model is better than the auto-regression in both cases, but logarithms are required only for high grade workers.

Question (5)

The judgements of three different levels of albedos by four observers at a level of illumination of 2 were given in Table **??**. The experiment was also performed, with the same observers, at an illumination of 1.2. The results are given below (McNemar, 1954, p. 321).

	Level of albedos		
Observer	0.07	0.14	0.26
1	11	24	60
2	22	26	44
3	16	22	55
4	20	32	82

(a) Set up the ANOVA table.

(b) Calculate the intra-class correlation and the differences in evaluation with level of albedos.

(c) Compare your results with those for the lower level of illumination.

(d) In Section 5.3.2, we performed an analysis of variance with two explanatory variables. Could this be extended to study level of albedos and level of illumination simultaneously?

Answer

(a) The ANOVA table is

	SS	MSS	d.f.	F
Individual	415.0	138.3	3	
Albedos level	4131.5	2065.8	2	26.0
Residual	476.5	79.4	6	

(b) The variance estimates are $\hat{\sigma}^2 = 79.4$ and $\hat{\sigma}_c^2 = 14.7$, so that the estimate of the intraclass correlation is $\hat{\rho} = 0.16$. The parameter estimates for the albedos levels are $\hat{\beta}_1 = -17.25$, $\hat{\beta}_2 = -8.50$ and $\hat{\beta}_3 = 25.75$, with $\hat{\mu} = 34.50$.

(c) The intraclass correlation here is considerably smaller than that (0.32) obtained for the lower level. On the other hand, the parameter estimates for differences with albedos level are very similar.

(d) The extension to two explanatory variables is straight forward, although it involves considerable calculations. The ANOVA table is

	SS	MSS	d.f.	F
Individual	1302.8	434.3	3	
Albedos level	8039.1	4019.5	2	25.8
Illumination	204.2	204.2	1	1.3
Interaction	46.6	23.3	2	0.1
Residual	1201.2	80.1	6	
Residual+Inter.	1247.8	156.0	8	

As might be expected from the previous two analyses, neither the interaction nor the difference in illumination is necessary in the model. Neither P-value is significant at the 10% level. Note again that, since the interaction test was not significant, the interaction and the residual lines in the ANOVA table were added to compute the mean sum of squares used as denominator for testing the main effects. The parameter estimates for the albedos levels are now $\hat{\beta}_1 = -18.42$, $\hat{\beta}_2 = -6.54$ and $\hat{\beta}_3 = 24.96$, with $\hat{\mu} = 37.42$. They are in between those for the two illuminations separately.

Question (6)

The Panel Study of Income Dynamics carried out in the USA contains information on unemployment periods due to layoffs. The sample distinguishes two ways in which the unemployment spell could end: by being recalled to the same job or finding a new job. The results are given in the table on the following page (Han and Hausman, 1990). Data for which durations can end in more than one way are called 'competing risks'. Usually, strong assumptions have to be made in order to model them.

Week	New job	Recall	Censor	Week	New job	Recall	Censor
1	10	93	0	36	2	1	0
2	8	118	0	37	0	1	2
3	8	55	0	38	1	0	0
4	23	58	0	39	5	4	7
5	3	18	0	40	4	1	1
6	11	26	0	41	1	0	0
7	1	6	0	42	0	0	2
8	22	38	0	43	1	4	2
9	6	13	1	44	0	0	0
10	7	10	0	45	1	0	0
11	4	4	0	46	0	0	0
12	13	32	1	47	0	0	2
13	10	19	9	48	0	0	1
14	0	9	2	49	1	0	1
15	4	14	2	50	1	1	0
16	10	9	3	51	0	0	0
17	8	7	18	52	4	0	23
18	5	2	6	53	1	0	0
19	2	0	3	54	0	0	0
20	9	12	4	55	0	0	2
21	3	1	7	56	1	0	0
22	5	7	9	57	0	0	1
23	1	0	2	58	0	0	0
24	7	10	4	59	0	0	0
25	2	1	2	60	1	0	1
26	18	15	21	61	0	0	2
27	0	2	1	62	0	0	0
28	0	2	0	63	0	0	0
29	1	0	1	64	0	0	0
30	9	4	9	65	0	0	1
31	0	0	3	66	1	0	1
32	1	0	1	67	0	1	1
33	1	0	0	68	0	0	0
34	2	1	3	69	0	1	0
35	2	0	8	70	4	3	33

(a) Ignoring the reason for unemployment ending, plot the Kaplan–Meier survivor curve.

(b) One possible approach to modelling competing risks is to assume that all terminations except that currently of interest are forms of censoring.

i Plot the Kaplan–Meier curve for obtaining a new job, assuming that those recalled are censored (as well as those actually censored).

ii Does any simple logistic model fit these data well?

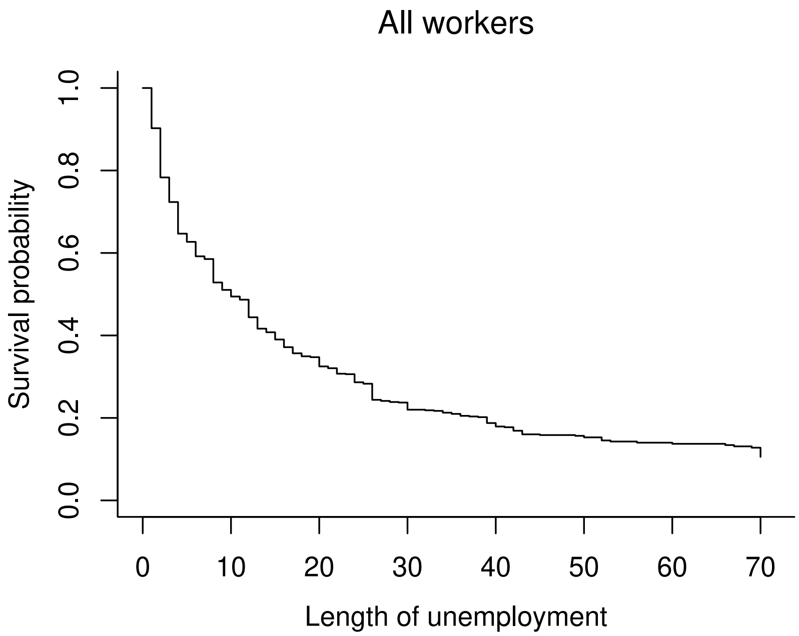

Figure 6.6: Kaplan-Meier curve for all termination of unemployment.

iii Use the same approach for recalls, assuming that finding a new job is censoring.

(c) Discuss possible drawbacks of such an approach to competing risks.

Answer

(a) The Kaplan-Meier curve ignoring the reason for unemployment ending is plotted in Figure **??**. Notice how the curve drops rapidly before levelling off. Some workers find a job rapidly, but then the rest have much more difficulty.

(b) The Kaplan-Meier curve for obtaining a new job is plotted in Figure **??**. This curve is relatively straight showing that workers do not necessarily find a new job quickly.

A logistic model with obtaining a new job depending on log time (AIC 115.0) fits better than one with it depending on time (136.6). The null model has an AIC of 136.9.

The Kaplan-Meier curve for being recalled is plotted in Figure **??**. This drops like the curve for all workers, showing that most recalls happen rather soon after unemployment starts.

A logistic model with being recalled depending on log time (AIC 109.1) also fits better than one with it depending on time (143.1). The null model has an AIC of 485.2.

(c) People who are censored may be very different from those finding a job so that it may not be wise to combine the two. There may also be a dependence between recall

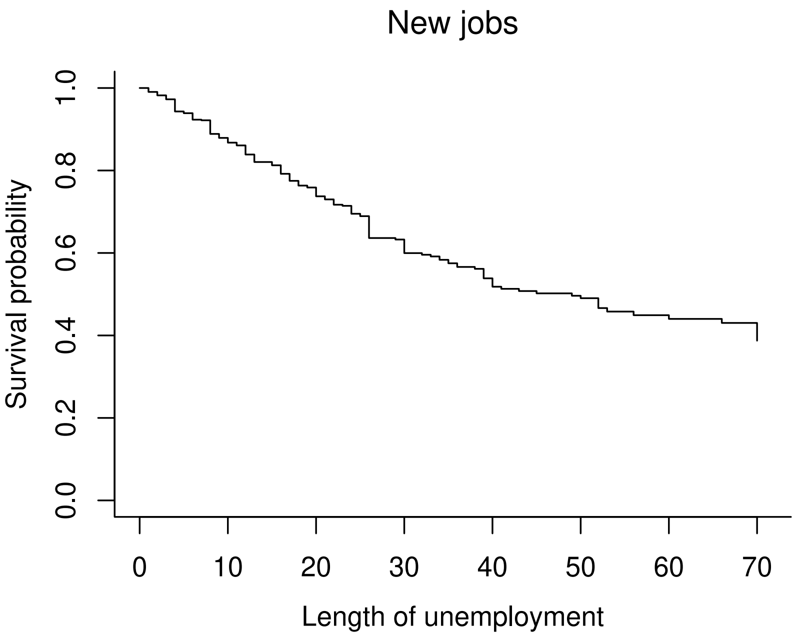

Figure 6.7: Kaplan-Meier curve for termination of unemployment by a new job.

and a new job. People who are recalled may also be those most likely to be able to find a new job.

Question (7)

Table **??** gave the survival over a ten year period of women with Stage II cancer of the cervix.

(a) Construct the life table for these data using both the Kaplan–Meier (binomial) and Aalen–Nelson (Poisson) methods.

(b) Women with Stage II cancer have a more advanced form of the disease than those with Stage I in Table **??**. Do the forms of their two survival curves support this fact?

(c) Compare the probabilities of censoring in the two stages.

Answer

(a) The life table for the women with Stage II cancer is given in Table **??**.

(b) The two survival curves are plotted in Figure **??**. As might be expected, the lower curve for Stage II indicates that these women survive less time.

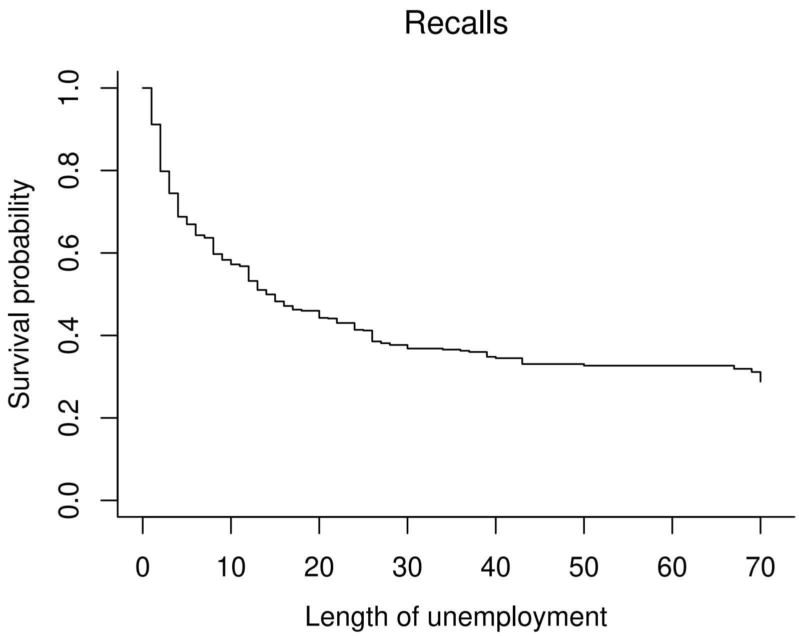

Figure 6.8: Kaplan-Meier curve for termination of unemployment by recall.

(c) We can set up a $10 \times 2 \times 2$ contingency table containing the numbers censored and not censored, out of those at risk, each year. The logistic regression gives a deviance of 78.90 on 19 degrees of freedom, clearly indicating that censoring depends either on the year or the stage or both. The parameter estimates for year are $\hat{\alpha} = (-2.19, -1.51, -0.54, -0.46, -0.28, 0.19, 0.16, 0.09, 0.50, 0.78, 1.07)$. They show that censoring generally increased over the years of the study. The estimate for difference with stage is $\hat{\beta}_1 = 0.22$, indicating more censoring in the Stage I group.

Question (8)

Table **??** gave the lengths of marriage before divorce in Liège. Notice that there is no censoring in these data.

(a) Plot the Kaplan–Meier curve for these data.

(b) Reconstruct the data as a contingency table and compare the fits of any appropriate logistic models.

(c) Discuss the complications in interpreting such a graph and models, given the special way in which the data were collected.

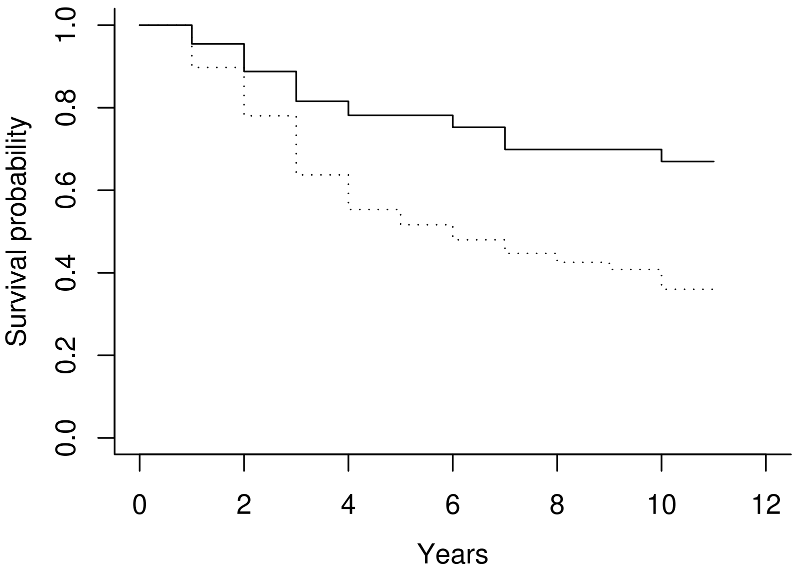

Figure 6.9: Survival curve for the Stage 2 cancer data. (from Clayton and Hills, 1993, p. 32)

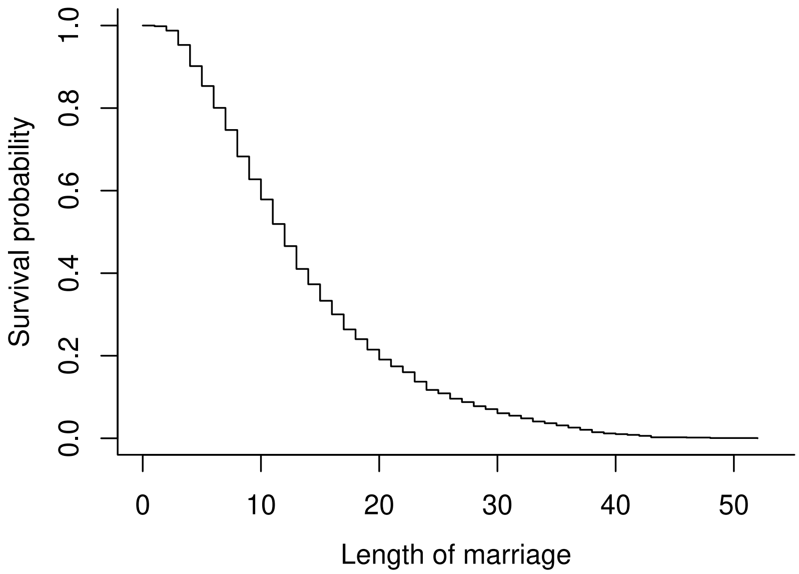

Figure 6.10: Kaplan-Meier curve for the divorce data.

Answer

(a) The Kaplan-Meier curve is plotted in Figure **??**.

(b) As with the unemployment data in Exercise (**??**.**??**), the model with the logarithm of time (AIC 160.2) fits better than that with time (344.1). The AIC of the null model is 676.8.

(c) It is very difficult to specify a larger population to which such data might be generalised. The couples were married spread over 50 different years before the time of the study. Because no information is available about couples who did not divorce, no conclusions can be made about length of marriage in general, only about length of marriage of couples divorcing in Liège in that year.

Table 6.3: Survival over a ten year period of women with two stages of cancer of the cervix. (Clayton and Hills, 1993, p. 32)

		Stage 1			Stage 2	
Years	Number	Deaths	Censored	Number	Deaths	Censored
1	110	5	5	234	24	3
2	100	7	7	207	27	11
3	86	7	7	169	31	9
4	72	3	8	129	17	7
5	61	0	7	105	7	13
6	54	2	10	85	6	6
7	42	3	6	73	5	6
8	33	0	5	62	3	10
9	28	0	4	49	2	13
10	24	1	8	34	4	6

Table 6.4: Estimated intensities and cumulative survival probabilities for the Stage 2 data. (from Clayton and Hills, 1993, p. 32)

		Binomial	Poisson
Years	Intensity	survival	survival
1	0.103	0.897	0.903
2	0.130	0.780	0.792
3	0.183	0.637	0.659
4	0.138	0.553	0.578
5	0.067	0.516	0.541
6	0.072	0.480	0.504
7	0.068	0.447	0.470
8	0.048	0.425	0.448
9	0.041	0.408	0.430
10	0.118	0.360	0.383