CHAPTER 1

INTRODUCTION

This chapter introduces basic statistical ideas and terminology in what the author hopes is a suitably concise fashion. Many readers will be able to turn to Chapter 2 without further ado!

1.1 WHAT ARE CATEGORICAL DATA?

Categorical data are the observed values of variables such as the color of a book, a person's religion, gender, political preference, social class, etc. In short, any variable other than a *continuous variable* (such as length, weight, time, distance, etc.).

If the categories have no obvious order (e.g., Red, Yellow, White, Blue) then the variable is described as a *nominal variable*. If the categories have an obvious order (e.g., Small, Medium, Large) then the variable is described as an *ordinal variable*. In the latter case the categories may relate to an underlying continuous variable where the precise value is unrecorded, or where it simplifies matters to replace the measurement by the relevant category. For example, while an individual's age may be known, it may suffice to record it as belonging to one of the categories "Under 18," "Between 18 and 65," "Over 65."

If a variable has just two categories, then it is a *binary variable* and whether or not the categories are ordered has no effect on the ensuing analysis.

Categorical Data Analysis by Example, First Edition. Graham J. G. Upton.

1.2 A TYPICAL DATA SET

The basic data with which we are concerned are *counts*, also called *frequencies*. Such data occur naturally when we summarize the answers to questions in a survey such as that in Table 1.1.

TABLE 1.1 Hypothetical sports preference survey

Sports preference questionnaire
(A) Are you:- Male ☐ Female ☐?
(B) Are you:- Aged 45 or under ☐ Aged over 45 ☐?
(C) Do you:- Prefer golf to tennis ☐ Prefer tennis to golf ☐?

The people answering this (fictitious) survey will be classified by each of the three characteristics: gender, age, and sport preference. Suppose that the 400 replies were as given in Table 1.2 which shows that males prefer golf to tennis (142 out of 194 is 73%) whereas females prefer tennis to golf (161 out of 206 is 78%). However, there is a lot of other information available. For example:

- There are more replies from females than males.
- There are more tennis lovers than golf lovers.
- Amongst males, the proportion preferring golf to tennis is greater amongst those aged over 45 (78/102 is 76%) than those aged 45 or under (64/92 is 70%).

This book is concerned with models that can reveal all of these subtleties simultaneously.

TABLE 1.2 Results of sports preference survey

Category of response	Frequency
Male, aged 45 or under, prefers golf to tennis	64
Male, aged 45 or under, prefers tennis to golf	28
Male, aged over 45, prefers golf to tennis	78
Male, aged over 45, prefers tennis to golf	24
Female, aged 45 or under, prefers golf to tennis	22
Female, aged 45 or under, prefers tennis to golf	86
Female, aged over 45, prefers golf to tennis	23
Female, aged over 45, prefers tennis to golf	75

1.3 VISUALIZATION AND CROSS-TABULATION

While Table 1.2 certainly summarizes the results, it does so in a clumsily long-winded fashion. We need a more succinct alternative, which is provided in Table 1.3.

TABLE 1.3 Presentation of survey results by gender

	Male				Female		
Sport	45 and under	Over 45	Total	Sport	45 and under	Over 45	Total
Tennis	28	24	52	Tennis	86	75	161
Golf	64	78	142	Golf	22	23	45
Total	92	102	194	Total	108	98	206

A table of this type is referred to as a *contingency table*—in this case it is (in effect) a three-dimensional contingency table. The locations in the body of the table are referred to as the *cells* of the table. Note that the table can be presented in several different ways. One alternative is Table 1.4.

In this example, the problem is that the page of a book is two-dimensional, whereas, with its three classifying variables, the data set is essentially three-dimensional, as Figure 1.1 indicates. Each face of the diagram contains information about the 2×2 category combinations for two variables for some particular category of the third variable.

With a small table and just three variables, a diagram is feasible, as Figure 1.1 illustrates. In general, however, there will be too many variables and too many categories for this to be a useful approach.

TABLE 1.4 Presentation of survey results by sport preference

	Prefers tennis				Prefers golf		
Gender	45 and under	Over 45	Total	Gender	45 and under	Over 45	Total
Female	86	75	161	Female	22	23	45
Male	28	24	52	Male	64	78	142
Total	114	99	213	Total	86	101	187

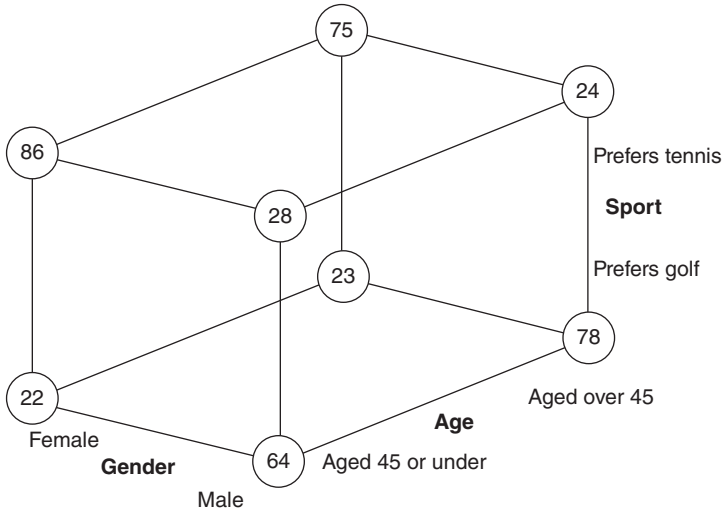

FIGURE 1.1 Illustration of results of sports preference survey.

1.4 SAMPLES, POPULATIONS, AND RANDOM VARIATION

Suppose we repeat the survey of sport preferences, interviewing a second group of 100 individuals and obtaining the results summarized in Table 1.5.

As one would expect, the results are very similar to those from the first survey, but they are not identical. All the principal characteristics (for example, the preference of females for tennis and males for golf) are again present, but there are slight variations because these are the replies from a different set of people. Each person has individual reasons for their reply and we cannot possibly expect to perfectly predict any individual reply since there can be thousands of contributing factors influencing a person's preference. Instead we attribute the differences to *random variation*.

TABLE 1.5 The results of a second survey

	Prefers tennis				Prefers golf		
Gender	45 and under	Over 45	Total	Gender	45 and under	Over 45	Total
Female	81	76	157	Female	16	24	40
Male	26	34	60	Male	62	81	143
Total	107	110	217	Total	78	105	183

Of course, if one survey was of spectators leaving a grand slam tennis tournament, whilst the second survey was of spectators at an open golf tournament, then the results would be very different! These would be *samples* from very different *populations*. Both samples may give entirely fair results for their own specialized populations, with the differences in the sample results reflecting the differences in the populations.

Our purpose in this book is to find succinct models that adequately describe the populations from which samples like these have been drawn. An effective model will use relatively few parameters to describe a much larger group of counts.

1.5 PROPORTION, PROBABILITY, AND CONDITIONAL PROBABILITY

Between them, Tables 1.4 and 1.5 summarized the sporting preferences of 800 individuals. The information was collected one individual at a time, so it would have been possible to keep track of the counts in the eight categories as they accumulated. The results might have been as shown in Table 1.6.

As the sample size increases, so the observed proportions, which are initially very variable, becomes less variable. Each proportion slowly converges on its limiting value, the *population probability*. The difference between columns three and five is that the former is converging on the probability of randomly selecting a particular type of individual from the whole population while the latter is converging on the *conditional probability* of selecting the individual from the relevant subpopulation (males aged over 40).

TABLE 1.6 The accumulating results from the two surveys

Sample size	Number of males over 40 who prefer golf	Proportion of sample that are males aged over 40 and prefer golf	Number of males	Proportion of males aged over 40 who prefer golf
10	3	0.300	6	0.500
20	5	0.250	11	0.455
50	8	0.160	25	0.320
100	22	0.220	51	0.431
200	41	0.205	98	0.418
400	78	0.195	194	0.402
800	159	0.199	397	0.401

1.6 PROBABILITY DISTRIBUTIONS

In this section, we very briefly introduce the distributions that are directly relevant to the remainder of the book. A variable is described as being a *discrete variable* if it can only take one of a finite set of values. The probability of any particular value is given by the *probability function*, P.

By contrast, a *continuous variable* can take any value in one or more possible ranges. For a continuous random variable the probability of a value in the interval (a, b) is given by integration of a function f (the so-called *probability density function*) over that interval.

1.6.1 The Binomial Distribution

The binomial distribution is a discrete distribution that is relevant when a variable has just two categories (e.g., Male and Female). If the probability of a randomly chosen individual has probability p of being male, then the probability that a random sample of n individuals contains r males is given by

$$\mathrm{P}(r) = \begin{cases} \binom{n}{r} p^r (1-p)^{n-r} & r = 0, 1, \ldots, n, \\ 0 & \text{otherwise}, \end{cases} \tag{1.1}$$

where

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

and

$$r! = r \times (r-1) \times (r-2) \times \cdots \times 2 \times 1.$$

A random variable having such a distribution has *mean* (the average value) np and *variance* (the usual measure of variability) $np(1-p)$. When p is very small and n is large—which is often the case in the context of contingency tables—then the distribution will be closely approximated by a Poisson distribution (Section 1.6.3) with the same mean. When n is large, a normal distribution (Section 1.6.4) also provides a good approximation.

This distribution underlies the logistic regression models discussed in Chapters 7–9.

1.6.2 The Multinomial Distribution

This is the extension of the binomial to the case where there are more than two categories. Suppose, for example, that a mail delivery company classifies packages as being either Small, Medium, and Large, with the proportions falling in these classes being p, q, and $1 - p - q$, respectively. The probability that a random sample of n packages includes r Small packages, s Medium packages, and $(n - r - s)$ Large packages is

$$\frac{n!}{r!s!(n-r-s)!}p^r q^s(1-p-q)^{n-r-s} \quad \text{where} \quad 0 \le r \le n; \quad 0 \le s \le (n-r).$$

This distribution underlies the models discussed in Chapter 10.

1.6.3 The Poisson Distribution

Suppose that the probability of an individual having a particular characteristic is p, independently, for each of a large number of individuals. In a random sample of n individuals, the probability that exactly r will have the characteristic, is given by Equation (1.1). However, if p (or $1 - p$) is small and n is large, then that binomial probability is well approximated by

$$\mathrm{P}(r) = \begin{cases} \dfrac{\mu^r}{r!}\mathrm{e}^{-\mu} & r = 0, 1, \ldots, \\ 0 & \text{otherwise,} \end{cases} \tag{1.2}$$

where e is the *exponential function* (=2.71828...) and $\mu = np$. A random variable with distribution given by Equation (1.2) is said to have a Poisson distribution with *parameter* (a value determining the shape of the distribution) μ. Such a random variable has both mean and variance equal to μ.

This distribution underlies the log-linear models discussed in Chapters 11–16.

1.6.4 The Normal Distribution

The normal distribution (known by engineers as the *Gaussian distribution*) is the most familiar example of a continuous distribution.

If X is a normal random variable with mean μ and variance σ^2, then X has probability density function given by

$$\mathrm{f}(x) = \frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{-(x-\mu)^2/2\sigma^2}. \tag{1.3}$$

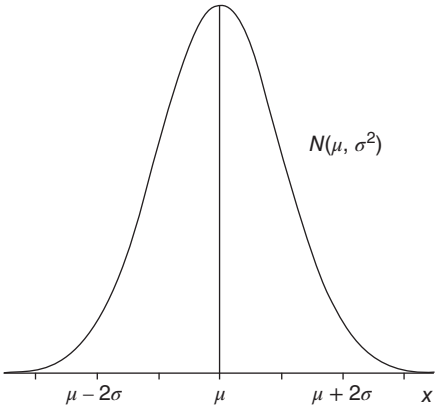

FIGURE 1.2 A normal distribution, with mean μ and variance σ^2.

The density function is illustrated in Figure 1.2. In the case where $\mu = 0$ and $\sigma^2 = 1$, the distribution is referred to as the *standard normal distribution*. Any tables of the normal distribution will be referring to this distribution.

Figure 1.2 shows that most (actually, about 95%) of observations on a random variable lie within about two *standard deviations* (actually 1.96σ) of the mean, with only about three observations in a thousand having values that differ by more than three standard deviations from the mean. The *standard deviation* is the square root of the variance.

1.6.4.1 The Central Limit Theorem

An informal statement of this theorem is

> A random variable that can be expressed as the sum of a large number of "component" variables which are independent of one another, but all have the same distribution, will have an approximate normal distribution.

The theorem goes a long way to explaining why the normal distribution is so frequently found, and why it can be used as an approximation to other distributions.

1.6.5 The Chi-Squared (χ^2) Distribution

A chi-squared distribution is a continuous distribution with a single parameter known as the *degrees of freedom* (often abbreviated as d.f.). Denoting the value of this parameter by ν, we write that a random variable has a χ^2_ν-distribution. The χ^2 distribution is related to the normal distribution since, if Z has a standard normal distribution, then Z^2 has a χ^2_1-distribution.

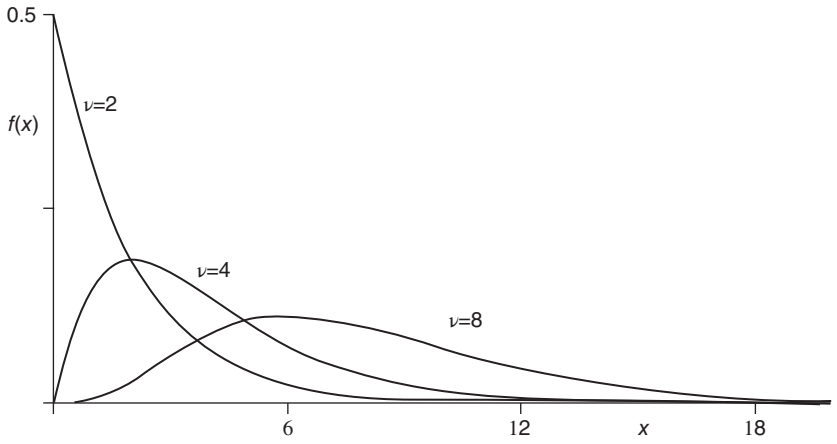

FIGURE 1.3 Chi-squared distributions with 2, 4, and 8 degrees of freedom.

Figure 1.3 gives an idea of what the probability density functions of chi-squared distributions look like. For small values of ν the distribution is notably skewed (for $\nu > 2$, the mode is at $\nu - 2$). A chi-squared random variable has mean ν and variance 2ν.

A very useful property of chi-squared random variables is their additivity: if U and V are independent random variables having, respectively χ^2_u- and χ^2_v-distributions, then their sum, $U + V$, has a χ^2_{u+v} distribution. This is known as the *additive* property of χ^2 distributions.

Perhaps more importantly, if W has a χ^2_w-distribution then it will always be possible to find w independent random variables (W_1, W_2,..., W_w) for which $W = W_1 + W_2 + \cdots + W_w$, with each of W_1, W_2,..., W_w having χ^2_1-distributions. We will make considerable use of this type of result in the analysis of contingency tables.

1.7 *THE LIKELIHOOD

Suppose that n observations, x_1, x_2,..., x_n, are taken on the random variable, X. The likelihood, L, is the product of the corresponding probability functions (in the case of a discrete distribution) or probability density functions (in the case of a continuous distribution):

$$\mathrm{L} = \mathrm{P}(x_1) \times \mathrm{P}(x_2) \times \cdots \times \mathrm{P}(x_n) \quad \text{or} \quad \mathrm{L} = \mathrm{f}(x_1) \times \mathrm{f}(x_2) \times \cdots \times \mathrm{f}(x_n) \quad (1.4)$$

In either case the likelihood is proportional to the probability that a future set of n observations have precisely the values observed in the current set. In most

cases, while the form of the distribution of X may be known (e.g., binomial, Poisson), the value of that distribution's parameter (e.g., p, μ) will not be known. Since the observed values have indeed occurred, a logical choice for any unknown parameter would be that value that maximizes the probability of reoccurrence of $x_1, x_2, \ldots, x_n$; this is the principle of *maximum likelihood*.

Suppose that there are r distinct values ($v_1, v_2, \ldots, v_r$) amongst the n observations, with the value v_i occurring on f_i occasions (so $\sum f_i = n$). The distribution that maximizes L is the discrete distribution that precisely describes what has been observed:

$$\mathrm{P}(X = x) = \begin{cases} f_i/n & \text{for } x = v_i, \quad i = 1, 2, \ldots, r, \\ 0 & \text{otherwise.} \end{cases} \tag{1.5}$$