1

Introduction

Statistics concerns what can be learned from data. Applied statistics comprises a body of methods for data collection and analysis across the whole range of science, and in areas such as engineering, medicine, business, and law — wherever variable data must be summarized, or used to test or confirm theories, or to inform decisions. Theoretical statistics underpins this by providing a framework for understanding the properties and scope of methods used in applications.

Statistical ideas may be expressed most precisely and economically in mathematical terms, but contact with data and with scientific reasoning has given statistics a distinctive outlook. Whereas mathematics is often judged by its elegance and generality, many statistical developments arise as a result of concrete questions posed by investigators and data that they hope will provide answers, and elegant and general solutions are not always available. The huge variety of such problems makes it hard to develop a single over-arching theory, but nevertheless common strands appear. Uniting them is the idea of a *statistical model*.

The key feature of a statistical model is that variability is represented using probability distributions, which form the building-blocks from which the model is constructed. Typically it must accommodate both random and systematic variation. The randomness inherent in the probability distribution accounts for apparently haphazard scatter in the data, and systematic pattern is supposed to be generated by structure in the model. The art of modelling lies in finding a balance that enables the questions at hand to be answered or new ones posed. The complexity of the model will depend on the problem at hand and the answer required, so different models and analyses may be appropriate for a single set of data.

Charles Robert Darwin (1809–1882) was rich enough not to have to earn his living. His reading and studies at Edinburgh and Cambridge exposed him to contemporary scientific ideas, and prepared him for the voyage of the Beagle (1831–1836), which formed the basis of his life's work as a naturalist — at one point he spent 8 years dissecting and classifying barnacles. He wrote numerous books including *The Origin of Species*, in which he laid out the theory of evolution by natural selection. Although his proposed mechanism for natural variation was never accepted, his ideas led to the biggest intellectual revolution of the 19th century, with repercussions that continue today. Ironically, his own family was in-bred and his health poor. See Desmond and Moore (1991).

Examples

Example 1.1 (Maize data) Charles Darwin collected data over a period of years on the heights of *Zea mays* plants. The plants were descended from the same parents and planted at the same time. Half of the plants were self-fertilized, and half were cross-fertilized, and the purpose of the experiment was to compare their heights. To

	Height (eighths of an inch)		
Pot	Crossed	Self-fertilized	Difference
I	188	139	49
	96	163	−67
	168	160	8
II	176	160	16
	153	147	6
	172	149	23
III	177	149	28
	163	122	41
	146	132	14
	173	144	29
	186	130	56
IV	168	144	24
	177	102	75
	184	124	60
	96	144	−48

Table 1.1 Heights of young *Zea mays* plants, recorded by Charles Darwin (Fisher, 1935a, p. 30).

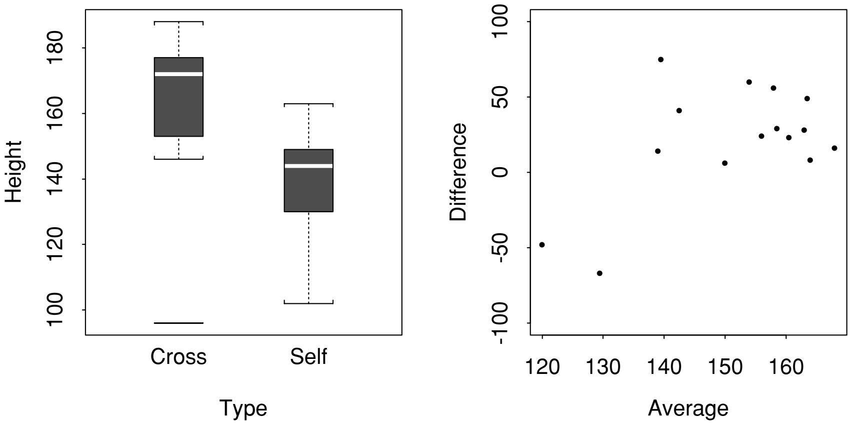

Figure 1.1 Summary plots for Darwin's *Zea mays* data. The left panel compares the heights for the two different types of fertilization. The right panel shows the difference for each pair plotted against the pair average.

this end Darwin planted them in pairs in different pots. Table 1.1 gives the resulting heights. All but two of the differences between pairs in the fourth column of the table are positive, which suggests that cross-fertilized plants are taller than self-fertilized ones.

This impression is confirmed by the left-hand panel of Figure 1.1, which summarizes the data in Table 1.1 in terms of a *boxplot*. The white line in the centre of each box shows the median or middle observation, the ends of each box show the observations roughly one-quarter of the way in from each end, and the bars attached to the box by the dotted lines show the maximum and minimum, provided they are not too extreme.

Cross-fertilized plants seem generally higher than self-fertilized ones. Overlaid on this systematic variation, there seems to be variation that might be ascribed to chance: not all the plants within each group have the same height. It might be possible,

Cambridge University Press
0521773393 - Statistical Models
A. C. Davison
Excerpt
More information

and for some purposes even desirable, to construct a mechanistic model for plant growth that could explain all the variation in such data. This would take into account genetic variation, soil and moisture conditions, ventilation, lighting, and so forth, through a vast system of equations requiring numerical solution. For most purposes, however, a deterministic model of this sort is quite unnecessary, and it is simpler and more useful to express variability in terms of probability distributions.

Francis Galton (1822–1911) was a cousin of Darwin from the same wealthy background. He explored in Africa before turning to scientific work, in which he showed a strong desire to quantify things. He was one of the first to understand the implications of evolution for *homo sapiens*, he invented the term regression and contributed to statistics as a by-product of his belief in the improvement of society via eugenics. See Stigler (1986).

Ronald Aylmer Fisher (1890–1962) was born in London and educated there and at Cambridge, where he had his first exposure to Mendelian genetics and the biometric movement. After obtaining the exact distributions of the *t* statistic and the correlation coefficient, but also having begun a life-long endeavour to give a Mendelian basis for Darwin's evolutionary theory, he moved in 1919 to Rothamsted Experimental Station, where he built the theoretical foundations of modern statistics, making fundamental contributions to likelihood inference, analysis of variance, randomization and the design of experiments. He wrote highly influential books on statistics and on genetics. He later held posts at University College London and Cambridge, and died in Adelaide. See Fisher Box (1978).

If the spread of heights within each group is modelled by random variability, the same cause will also generate variation between groups. This occurred to Darwin, who asked his cousin, Francis Galton, whether the difference in heights between the types of plants was too large to have occurred by chance, and was in fact due to the effect of fertilization. If so, he wanted to estimate the average height increase. Galton proposed an analysis based essentially on the following model. The height of a self-fertilized plant is taken to be

$$Y = \mu + \sigma\varepsilon, \tag{1.1}$$

where μ and σ are fixed unknown quantities called *parameters*, and ε is a random variable with mean zero and unit variance. Thus the mean of Y is μ and its variance is σ^2. The height of a cross-fertilized plant is taken to be

$$X = \mu + \eta + \sigma\varepsilon, \tag{1.2}$$

where η is another unknown parameter. The mean height of a cross-fertilized plant is $\mu + \eta$ and its variance is σ^2. In (1.1) and (1.2) variation within the groups is accounted for by the randomness of ε, whereas variation between groups is modelled deterministically by the difference between the means of Y and X. Under this model the questions posed by Darwin amount to:

- is η non-zero?
- Can we estimate η and state the uncertainty of our estimate?

Galton's analysis proceeded as if the observations from the self-fertilized plants, $Y_1, \dots, Y_{15}$, were independent and identically distributed according to (1.1), and those from the cross-fertilized plants, $X_1, \dots, X_{15}$, were independent and identically distributed according to (1.2). If so, it is natural to estimate the group means by $\overline{Y} = (Y_1 + \cdots + Y_{15})/15$ and $\overline{X} = (X_1 + \cdots + X_{15})/15$, and to compare $\overline{Y}$ and $\overline{X}$. In fact Galton proposed another analysis which we do not pursue.

In discussing this experiment many years later, R. A. Fisher pointed out that the model based on (1.1) and (1.2) is inappropriate. In order to minimize differences in humidity, growing conditions, and lighting, Darwin had taken the trouble to plant the seeds in pairs in the same pots. Comparison of different pairs would therefore involve these differences, which are not of interest, whereas comparisons within pairs would depend only on the type of fertilization. A model for this writes

$$Y_j = \mu_j + \sigma\varepsilon_{1j}, \quad X_j = \mu_j + \eta + \sigma\varepsilon_{2j}, \quad j = 1, \dots, 15. \tag{1.3}$$

The parameter μ_j represents the effects of the planting conditions for the jth pair, and the ε_{gj} are taken to be independent random variables with mean zero and unit

	Stress (N/mm^2)					
	950	900	850	800	750	700
	225	216	324	627	3402	12510+
	171	162	321	1051	9417	12505+
	198	153	432	1434	1802	3027
	189	216	252	2020	4326	12505+
	189	225	279	525	11520+	6253
	135	216	414	402	7152	8011
	162	306	396	463	2969	7795
	135	225	379	431	3012	11604+
	117	243	351	365	1550	11604+
	162	189	333	715	11211	12470+
$\overline{y}$	168	215	348	803	5636	9828
s	33	43	58	544	3864	3355

Table 1.2 Failure times (in units of 10^3 cycles) of springs at cycles of repeated loading under the given stress (Cox and Oakes, 1984, p. 8). + indicates that an observation is right-censored. The average and estimated standard deviation for each level of stress are $\overline{y}$ and s.

variance. The μ_j could be eliminated by basing the analysis on the $X_j - Y_j$, which have mean η and variance $2\sigma^2$.

The right panel of Figure 1.1 shows a *scatterplot* of pair differences $x_j - y_j$ against pair averages $(y_j + x_j)/2$. The two negative differences correspond to the pairs with the lowest averages. The averages vary widely, and it seems wise to allow for this by analyzing the differences, as Fisher suggested. ■

Both models in Example 1.1 summarize the effect of interest, namely the mean difference in heights of the plants, in terms of a fixed but unknown parameter. Other aspects of secondary interest, such as the mean height of self-fertilized plants, are also summarized by the parameters μ and σ of (1.1) and (1.2), and $\mu_1, \ldots, \mu_{15}$ and σ of (1.3). But even if the values of all these parameters were known, the distributions of the heights would still not be known completely, because the distribution of ε has not been fully specified. Such a model is called *nonparametric*. If we were willing to assume that ε has a given distribution, then the distributions of Y and X would be completely specified once the parameters were known, giving a *parametric model*. Most of this book concerns such models.

The focus of interest in Example 1.1 is the relation between the height of a plant and something that can be controlled by the experimenter, namely whether it is self- or cross-fertilized. The essence of the model is to regard the height as random with a distribution that depends on the type of fertilization, which is fixed for each plant. The variable of primary interest, in this instance height, is called the *response*, and the variable on which it depends, the type of fertilization, is called an *explanatory variable* or a *covariate*. Many questions arising in data analysis involve the dependence of one or more variables on another or others, but virtually limitless complications can arise.

Example 1.2 (Spring failure data) In industrial experiments to assess their reliability, springs were subjected to cycles of repeated loading until they failed. The failure 'times', in units of 10^3 cycles of loading, are given in Table 1.2. There were 60 springs divided into groups of 10 at each of six different levels of stress.

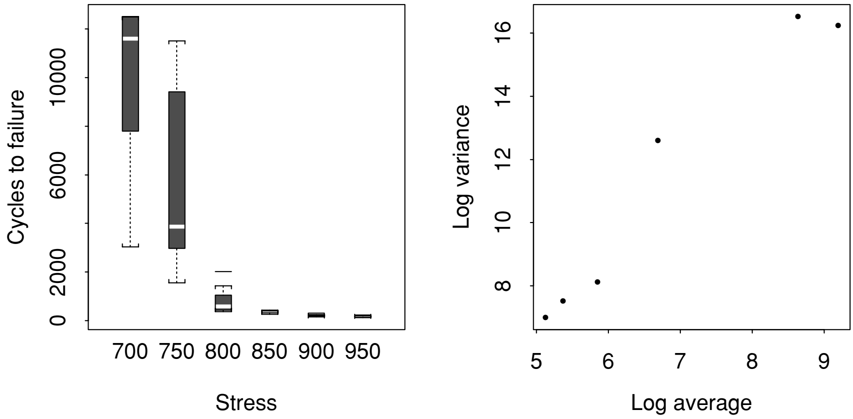

Figure 1.2 Failure times (in units of 10^3 cycles) of springs at cycles of repeated loading under the given stress. The left panel shows failure time boxplots for the different stresses. The right panel shows a rough linear relation between log average and log variance at the different stresses.

As stress decreases there is a rapid increase in the average number of cycles to failure, to the extent that at the lowest levels, where the failure time is longest, the experiment had to be stopped before all the springs had failed. The observations are *right-censored*: the recorded value is a lower bound for the number of cycles to failure that would have been observed had the experiment been continued to the bitter end. A right-censored observation is indicated as, say, 11520+, indicating that the failure time would be greater than 11520.

Let us represent the jth number of cycles to failure at the kth loading by y_{lj}, for $j = 1, \ldots, 10$ and $l = 1, \ldots, 6$. Table 1.2 shows the average failure time for each loading, $\overline{y}_{l\cdot} = 10^{-1} \sum_j y_{lj}$, and the sample standard deviation, s_l, where the sample variance is $s_l^2 = (10-1)^{-1} \sum_j (y_{lj} - \overline{y}_{l\cdot})^2$. The average and variance at the lowest stresses underestimate the true values, because of the censoring. The average and standard deviation decrease as stress increases.

The boxplots in the left panel of Figure 1.2 show that the cycles to failure at each stress have the marked pattern already described. The right panel shows the log variance, $\log s_l^2$, plotted against the log average, $\log \overline{y}_{l\cdot}$. It shows a linear pattern with slope approximately two, suggesting that variance is proportional to mean squared for these data.

Our inspection has revealed that:

(a) failure times are positive and range from 117–12510×10^3 or more cycles;
(b) there is strong dependence between the mean and variance;
(c) there is strong dependence of failure time on stress; and
(d) some observations are censored.

To proceed further, we would need to know how the data were gathered. Do systematic patterns, of which we have been told nothing, underlie the data? For example, were all 60 springs selected at random from a larger batch and then allocated to the different stresses at random? Or were the ten springs at 950 N/mm^2 selected from one batch, the ten springs at 900 N/mm^2 from another, and so on? If so, the apparent dependence on stress might be due to differences among batches. Were all measurements made with the same machine? If the answers to these and other such questions were unsatisfactory, we might suggest that better data be produced by performing another experiment designed to control the effects of different sources of variability.

Suppose instead that we are provisionally satisfied that we can treat observations at each loading as independent and identically distributed, and that the apparent dependence between cycles to failure and stress is not due to some other factor. With (a) and (b) in mind, we aim to represent the failure time at a given stress level by a random variable Y that takes continuous positive values and whose probability density function $f(y;\theta)$ keeps the ratio (mean)^2/variance constant. Clearly it is preferable if the same parametric form is used at each stress and the effect of changing stress enters only through θ. A simple model is that Y has exponential density

$$f(y;\theta) = \theta^{-1}\exp(-y/\theta), \quad y > 0, \theta > 0, \tag{1.4}$$

whose mean and variance are θ and θ^2, so that (mean)^2 = variance. We can express systematic variation in the density of Y in terms of stress, x, by

$$\theta = \frac{1}{\beta x}, \quad x > 0, \beta > 0, \tag{1.5}$$

though of course other forms of dependence are possible.

Equations (1.4) and (1.5) imply that when $x = 0$ the mean failure time is infinite, but it decreases to zero as stress x increases. Expression (1.4) represents the random component of the model, for a given value of θ, and (1.5) the systematic component, which determines how mean failure time θ depends on x. ■

In Examples 1.1 and 1.2 the response is continuous, and there is a single explanatory variable. But data with a discrete response or more than one explanatory variable often arise in practice.

Example 1.3 (Challenger data) The space shuttle Challenger exploded shortly after its launch on 28 January 1986, with a loss of seven lives. The subsequent US Presidential Commission concluded that the accident was caused by leakage of gas from one of the fuel-tanks. Rubber insulating rings, so-called ‘O-rings’, were not pliable enough after the overnight low temperature of 31°F, and did not plug the joint between the fuel in the tanks and the intense heat outside.

There are two types of joint, nozzle-joints and field-joints, each containing a primary O-ring and a secondary O-ring, together with putty that insulates both rings from the propellant gas. Table 1.3 gives the number of primary rings, r, out of the total $m = 6$ field-joints, that had experienced ‘thermal distress’ on previous flights. Thermal distress occurs when excessive heat pits the ring — ‘erosion’ — or when gases rush past the ring —- ‘blowby’. Blowby can occur in the short gap after ignition before an O-ring seals. It can also occur if the ring seals and then fails, perhaps because it has been eroded by the hot gas. Bench tests had suggested that one cause of blowby was that the O-rings lost their resilience at low temperatures. It was also suspected that pressure tests conducted before each launch holed the putty, making erosion of the rings more likely.

Table 1.3 O-ring thermal distress data. r is the number of field-joint O-rings showing thermal distress out of 6, for a launch at the given temperature (°F) and pressure (pounds per square inch) (Dalal *et al.*, 1989).

Flight	Date	Number of O-rings with thermal distress, r	Temperature (°F)	Pressure (psi)
			x_1	x_2
1	21/4/81	0	66	50
2	12/11/81	1	70	50
3	22/3/82	0	69	50
5	11/11/82	0	68	50
6	4/4/83	0	67	50
7	18/6/83	0	72	50
8	30/8/83	0	73	100
9	28/11/83	0	70	100
41-B	3/2/84	1	57	200
41-C	6/4/84	1	63	200
41-D	30/8/84	1	70	200
41-G	5/10/84	0	78	200
51-A	8/11/84	0	67	200
51-C	24/1/85	2	53	200
51-D	12/4/85	0	67	200
51-B	29/4/85	0	75	200
51-G	17/6/85	0	70	200
51-F	29/7/85	0	81	200
51-I	27/8/85	0	76	200
51-J	3/10/85	0	79	200
61-A	30/10/85	2	75	200
61-B	26/11/86	0	76	200
61-C	21/1/86	1	58	200
61-I	28/1/86	—	31	200

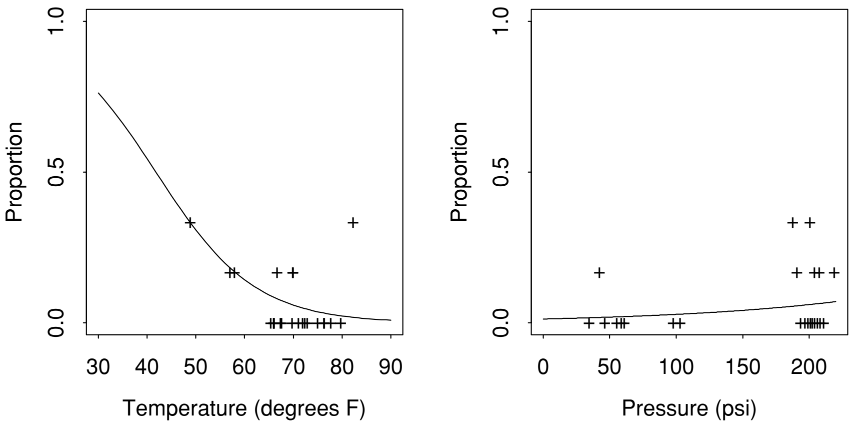

Figure 1.3 O-ring thermal distress data. The left panel shows the proportion of incidents as a function of joint temperature, and the right panel shows the corresponding plot against pressure. The x-values have been jittered to avoid overplotting multiple points. The solid lines show the fitted proportions of failures under a model described in Chapter 4.

Table 1.3 shows the temperatures x_1 and test pressures x_2 associated with thermal distress of the O-rings for flights before the disaster. The pattern becomes clearer when the proportion of failures, r/m, is plotted against temperature and pressure in Figure 1.3. As temperature decreases, r/m appears to increase. There is less pattern in the corresponding plot for pressure.

Table 1.4 Lung cancer deaths in British male physicians (Frome, 1983). The table gives man-years at risk/number of cases of lung cancer, cross-classified by years of smoking, taken to be age minus 20 years, and number of cigarettes smoked per day.

Years of smoking t	Daily cigarette consumption d						
	Nonsmokers	1–9	10–14	15–19	20–24	25–34	35+
15–19	10366/1	3121	3577	4317	5683	3042	670
20–24	8162	2937	3286/1	4214	6385/1	4050/1	1166
25–29	5969	2288	2546/1	3185	5483/1	4290/4	1482
30–34	4496	2015	2219/2	2560/4	4687/6	4268/9	1580/4
35–39	3512	1648/1	1826	1893	3646/5	3529/9	1336/6
40–44	2201	1310/2	1386/1	1334/2	2411/12	2424/11	924/10
45–49	1421	927	988/2	849/2	1567/9	1409/10	556/7
50–54	1121	710/3	684/4	470/2	857/7	663/5	255/4
55–59	826/2	606	449/3	280/5	416/7	284/3	104/1

For these data, the response variable takes one of the values 0, 1, . . . , 6, with fairly strong dependence on temperature and possibly weaker dependence on pressure. If we assume that at a given temperature and pressure, each of the six rings fails independently with equal probability, we can treat the number of failures R as binomial with denominator m and probability π,

$$\Pr(R = r) = \frac{m!}{r!(m-r)!}\pi^r(1-\pi)^{m-r}, \qquad r = 0, 1, \ldots, m, \; 0 < \pi < 1. \tag{1.6}$$

One possible relation between temperature x_1, pressure x_2, and the probability of failure is $\pi = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, where the parameters β_0, β_1, and β_2 must be derived from the data. This has the drawback of predicting probabilities outside the range [0, 1] for certain values of x_1 and x_2. It is more satisfactory to use a function such as

$$\pi = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)},$$

so $0 < \pi < 1$ wherever $\beta_0 + \beta_1 x_1 + \beta_2 x_2$ roams in the real line. It turns out that the function $e^u/(1 + e^u)$, the logistic distribution function, has an elegant connection to the binomial density, but any other continuous distribution function with domain the real line might be used.

The night before the Challenger was launched, there was a lengthy discussion about how the O-rings might behave at the low predicted launch temperature. One approach, which was not taken, would have been to try and predict how many O-rings might fail based on an estimated relationship between temperature and pressure. The lines in Figure 1.3 represent the estimated dependence of failure probability on x_1 and x_2, and show a high probability of failure at the actual launch temperature. When this is used as input to a probability model of how failures occur, the probability of catastrophic failure for a launch at 31°F is estimated to be as high as 0.16. To obtain this estimate involves extrapolation outside the available data, but there would have been little alternative in the circumstances of the launch. ■

Example 1.4 (Lung cancer data) Table 1.4 shows data on the lung cancer mortality of cigarette smokers among British male physicians. The table shows the man-years

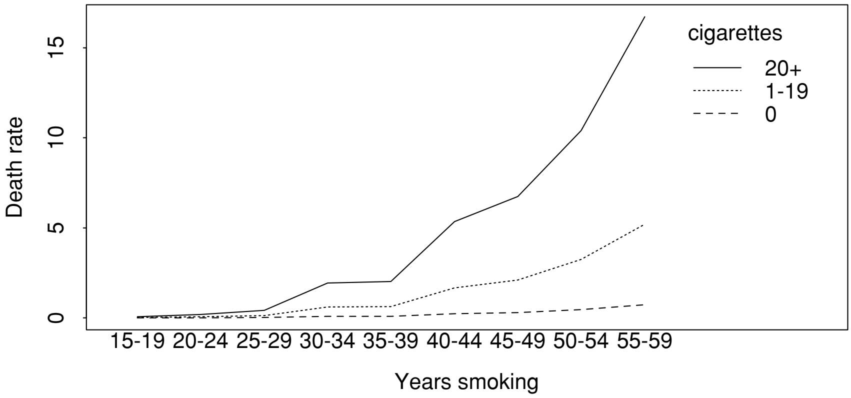

Figure 1.4 Lung cancer deaths in British male physicians. The figure shows the rate of deaths per 1000 man-years at risk, for each of three levels of daily cigarette consumption.

at risk and the number of cases with lung cancer, cross-classified by the number of years of smoking, taken to be age minus twenty years, and the number of cigarettes smoked daily. The man-years at risk in each category is the total period for which the individuals in that category were at risk of death.

As the eye moves from top left to the bottom right of the table, the figures suggest that death rate increases with increased total cigarette consumption. This is confirmed by Figure 1.4, which shows the death rate per 100,000 man-years at risk, grouped by three levels of cigarette consumption. Data for the first two groups show that death rate for smokers increases with cigarette consumption and with years of smoking. The only nonsmoker deaths are one in the age-group 35–39 and two in the age-group 75–79.

In this problem the aspect of primary interest is how death rate depends on cigarette consumption and smoking, and we treat the number of deaths in each category as the response. To build a model, we suppose that the death rate for those smoking d cigarettes per day after t years of smoking is $\lambda(d, t)$ deaths per man-year. Thus we may imagine deaths occurring at random in the total T man-years at risk in that category, at rate $\lambda(d, t)$. If deaths are independent point events in a continuum of length T, the number of deaths, Y, will have approximately a Poisson density with mean $T\lambda(d, t)$,

$$\Pr(Y = y) = \frac{\{T\lambda(d, t)\}^y}{y!} \exp\{-T\lambda(d, t)\}, \quad y = 0, 1, 2, \ldots. \tag{1.7}$$

One possible form for the mean deaths per man-year is

$$\lambda(d, t) = \beta_0 t^{\beta_1} \left(1 + \beta_2 d^{\beta_3}\right), \tag{1.8}$$

based on a deterministic argument and used in animal cancer mortality studies. In (1.8) there are four unknown parameters, and power-law dependence of death rate on exposure duration, t, and cigarette consumption, d. We expect that all the parameters β_r are positive. The background death-rate in the absence of smoking is given by $\beta_0 t^{\beta_1}$, the death-rate for nonsmokers. This represents the overall effect of other causes of lung cancer.

Expressions (1.7) and (1.8) give the random and systematic components for a simple model for the data, based on a blend of stochastic and deterministic arguments. An increasingly important development in statistics is the use of very complex models for real-world phenomena. Stochastic processes often provide the blocks with which such models are built. ■

There is an important difference between Example 1.4 and the previous examples. In Example 1.1, Darwin could decide which plants to cross and where to plant them, in Example 1.2 the springs could be allocated to different stresses by the experimenter, and in Example 1.3 the test pressure for field joints was determined by engineers. The engineers would have no control over the temperature at the proposed time of a launch, but they could decide whether or not to launch at a given temperature. In each case, the allocation of treatments could in principle be controlled, albeit to different extents. Such situations, called *controlled experiments*, often involve a random allocation of treatments — type of fertilization, level of stress or test pressure — to units — plants, springs, or flights. Strong conclusions can in principle be drawn when randomization is used — though it played no part in Examples 1.1 or 1.3, and we do not know about Example 1.2.

In Example 1.4, however, a new problem rears its head. There is no question of allocating a level of cigarette consumption over a given period to individuals — the practical difficulties would be insuperable, quite apart from ethical considerations. In common with many other epidemiological, medical, and environmental studies, the data are *observational*, and this limits what conclusions may be drawn. It might be postulated that propensities to smoking and to lung cancer were genetically related, causing the apparent dependence in Table 1.4. Then for an individual to stop smoking would not reduce their chance of contracting lung cancer. In such cases data of different types from different sources must be gathered and their messages carefully collated and interpreted in order to put together an unambiguous story.

Despite differences in interpretation, the use of probability models to summarize variability and express uncertainty is the basis of each example. It is the subject of this book.

Outline

The idea of treating data as outcomes of random variables has implications for how they should be treated. For example, graphical and numerical summaries of the observations will show variation, and it is important to understand its consequences. Chapter 2 is devoted to this. It deals with basic ideas such as parameters, statistics, and sampling variation, simple graphs and other summary quantities, and then turns to notions of convergence, which are essential for understanding variability in large samples and generating approximations for small ones. Many statistics are based on quantities such as the largest item in a sample, and order statistics are also discussed. The chapter finishes with an account of moments and cumulants.