Statistical Models

A. C. Davison

Department of Mathematics
Swiss Federal Institute of Technology
1015 Lausanne
Switzerland
Anthony.Davison@epfl.ch

April 12, 2003

Contents

Preface

A statistical model is a probability distribution constructed to enable inferences to be drawn or decisions made from data. This idea is the basis of most tools in the statistical workshop, in which it plays a central role by providing economical and insightful summaries of the information available.

This book is intended as an integrated modern account of statistical models covering the core topics for studies up to a masters degree in statistics. It can be used for a variety of courses at this level and for reference. After outlining basic notions, it contains a treatment of likelihood that includes non-regular cases and model selection, followed by sections on topics such as Markov processes, Markov random fields, point processes, censored and missing data, and estimating functions, as well as more standard material. Simulation is introduced early to give a feel for randomness, and later used for inference. There are major chapters on linear and nonlinear regression and on Bayesian ideas, the latter sketching modern computational techniques. Each chapter has a wide range of examples intended to show the interplay of subject-matter, mathematical, and computational considerations that makes statistical work so varied, so challenging, and so fascinating.

The target audience is senior undergraduate and graduate students, but the book should also be useful for others wanting an overview of modern statistics. The reader is assumed to have a good grasp of calculus and linear algebra, and to have followed a course in probability including joint and conditional densities, moment-generating functions, elementary notions of convergence and the central limit theorem, for example using Grimmett and Welsh (1986) or Stirzaker (1994). Measure is not required. Some sections involve a basic knowledge of stochastic processes, but they are intended to be as self-contained as possible. To have included full proofs of every statement would have made the book even longer and very tedious. Instead I have tried to give arguments for simple cases, and to indicate how results generalize. Readers in search of

mathematical rigour should see Knight (2000), Schervish (1995), Shao (1999), or van der Vaart (1998), amongst the many excellent books on mathematical statistics.

Solution of problems is an integral part of learning a mathematical subject. Most sections of the book finish with exercises that test or deepen knowledge of that section, and each chapter ends with problems which are generally broader or more demanding.

Real understanding of statistical methods comes from contact with data. Appendix 1 outlines practicals intended to give the reader this experience. The practicals themselves can be downloaded from

`http://statwww.epfl.ch/davison/SM/`

together with a library of functions and data to go with the book, and errata. The practicals are written in two dialects of the `S` language, for the freely available package `R` and for the commercial package `S-plus`, but it should not be hard for teachers to translate them for use with other packages.

Biographical sketches of some of the people mentioned in the text are given as sidenotes; the sources for many of these are Heyde and Seneta (2001) and

`http://www-groups.dcs.st-and.ac.uk/~history/`

Part of the work was performed while I was supported by an Advanced Research Fellowship from the UK Engineering and Physical Science Research Council. I am grateful to them and to my past and present employers for sabbatical leaves during which the book advanced. Many people have helped in various ways, for example by supplying data, examples, or figures, by commenting on the text, or by testing the problems. I thank Marc-Olivier Boldi, Alessandra Brazzale, Angelo Canty, Gorana Capkun, James Carpenter, Valérie Chavez, Stuart Coles, John Copas, Tom DiCiccio, Debbie Dupuis, David Firth, Christophe Girardet, David Hinkley, Wilfred Kendall, Diego Kuonen, Stephan Morgenthaler, Christophe Osinski, Brian Ripley, Gareth Roberts, Sylvain Sardy, Jamie Stafford, Trevor Sweeting, Valérie Ventura, Simon Wood, and various anonymous reviewers. Particular thanks go to Jean-Yves Le Boudec, Nancy Reid, and Alastair Young, who gave valuable comments on much of the book. David Tranah of Cambridge University Press displayed exemplary patience during the interminable wait for me to finish. Despite all their efforts, errors and obscurities doubtless remain. I take responsibility for this and would appreciate being told of them, in order to correct any future versions.

I dedicate this book to my long-suffering family, and particularly to Claire, without whose love and support the project would never have been finished.

Lausanne, January 2003

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.

Pot	Height (eighths of an inch)		Difference
	Crossed	Self-fertilized	
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).

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 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, 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 deter-

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).

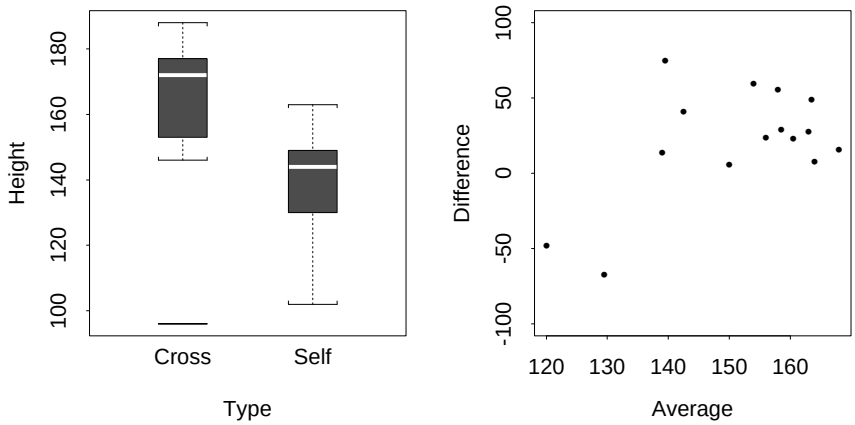

Figure 1.1 Summary plots for Darwin's *Zea mays* data. The left panel compares the heights for the two different types of fertilization. The right panel shows the difference for each pair plotted against the pair average.

ministic model of this sort is quite unnecessary, and it is simpler and more useful to express variability in terms of probability distributions.

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, $Y_1, \ldots, Y_{15}$, were independent and identically distributed according to (1.1),

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).

and those from the cross-fertilized plants, $X_1, \ldots, X_{15}$, 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

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

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 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. ■

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).

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

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.

	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

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.

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:

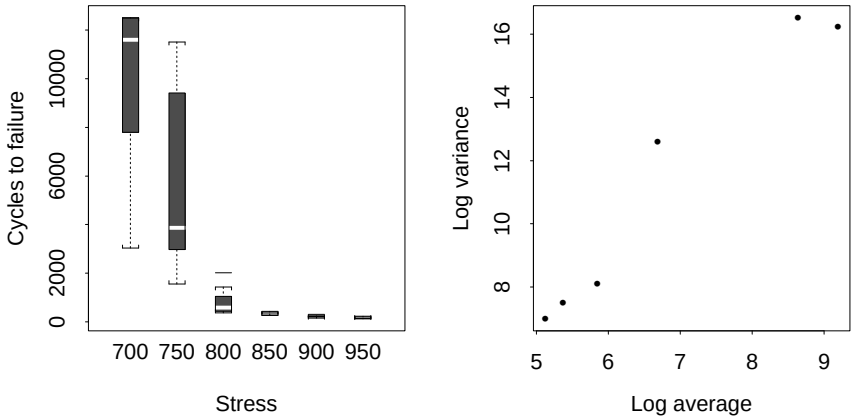

Figure 1.2 Failure times (in units of 10^3 cycles) of springs at cycles of repeated loading under the given stress. The left panel shows failure time boxplots for the different stresses. The right panel shows a rough linear relation between log average and log variance at the different stresses.

(a) failure times are positive and range from 117–12510$\times 10^3$ 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 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.

For these data, the response variable takes one of the values $0, 1, \ldots, 6$, 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

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) x_1	Pressure (psi) 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

six rings fails independently with equal probability, we can treat the number of failures R as binomial with denominator m and probability π,

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

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.

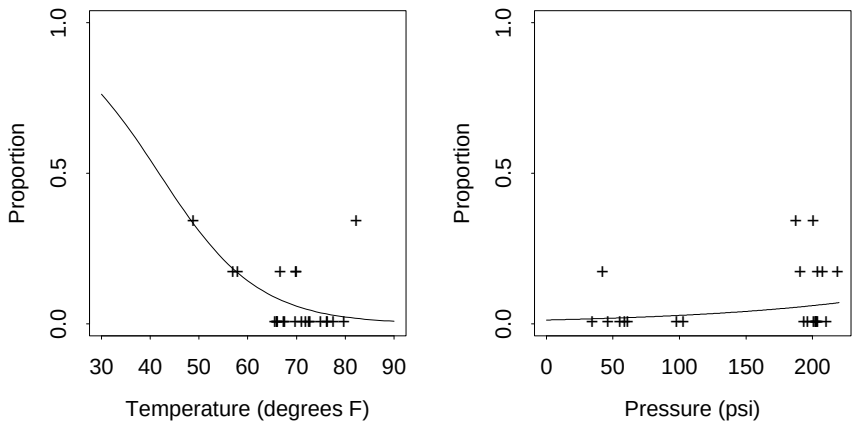

Figure 1.3 O-ring thermal distress data. The left panel shows the proportion of incidents as a function of joint temperature, and the right panel shows the corresponding plot against pressure. The x-values have been jittered to avoid overplotting multiple points. The solid lines show the fitted proportions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

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 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.

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

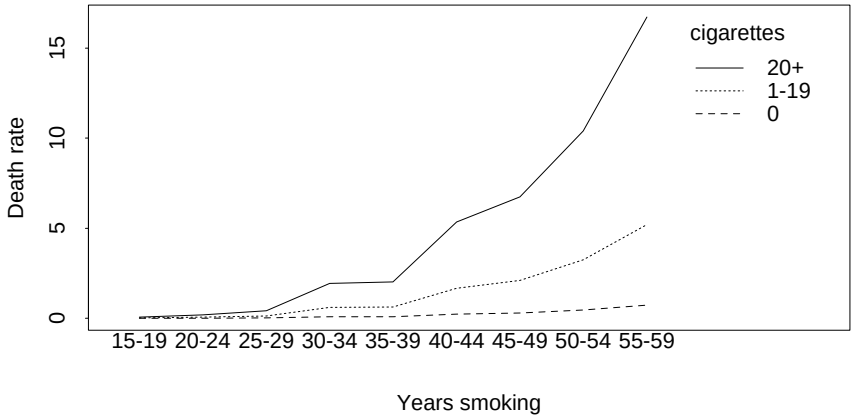

Figure 1.4 Lung cancer deaths in British male physicians. The figure shows the rate of deaths per 1000 man-years at risk, for each of three levels of daily cigarette consumption.

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.

Variation in observed data leads to uncertainty about the reality behind it. Uncertainty is a more complicated notion, because it entails considering what it is reasonable to infer from the data, and people differ in what they find reasonable. Chapter 3 explains one of the main approaches to expressing uncertainty, leading to the construction of confidence intervals via quantities known as pivots. In most cases these can only be approximate, but they are often exact for models based on the normal distribution, which are then described. The chapter ends with a brief account of Monte Carlo simulation, which is used both to appreciate variability and to assess uncertainty.

In some cases information about model parameters θ can be expressed as a density $\pi(\theta)$, separate from the data y. Then the prior uncertainty $\pi(\theta)$ may be updated to posterior uncertainty $\pi(\theta \mid y)$ using Bayes' theorem

$$\pi(\theta \mid y) = \frac{\pi(\theta) f(y \mid \theta)}{f(y)},$$

Thomas Bayes (1702–1761) was a nonconformist minister and also a mathematician. His theorem is contained in his *Essay towards solving a problem in the doctrine of chances*, found in his papers after his death and published in 1764.

which converts the conditional density $f(y \mid \theta)$ of observing data y, given that the true parameter is θ, into a conditional density for θ, given that y has been observed. This Bayesian approach to inference is attractive and conceptually simple, and modern computing techniques make it feasible to apply it to many complex models. However many statisticians do not agree that prior knowledge can or indeed should always be expressed as a prior density, and believe that information in the data should be kept separate from prior beliefs, preferring to base inference on the second term $f(y \mid \theta)$ in the numerator of Bayes' theorem, known as the likelihood.

Likelihood is a central idea for parametric models, and it and its ramifications are described in Chapter 4. Definitions of likelihood, the maximum likelihood estimator and information are followed by a discussion of inference based on maximum likelihood estimates and likelihood ratio statistics. The chapter ends with brief accounts of non-regular models and model selection.

Chapters 5 and 6 describe some particular classes of models. Accounts are

given of the simplest form of linear model, of exponential family and group transformation models, of models for survival and missing data, and of those with more complex dependence structures such as Markov chains, Markov random fields, point processes, and the multivariate normal distribution.

Chapter 7 discusses more traditional topics of mathematical statistics, with a more general treatment of point and interval estimation and testing than in the previous chapters. It also includes an account of estimating functions, which are needed subsequently.

Regression models describe how a response variable, treated as random, depends on explanatory variables, treated as fixed. The vast majority of statistical modelling involves some form of regression, and three chapters of the book are devoted to it. Chapter 8 describes the linear model, including its basic properties, analysis of variance, model building, and variable selection. Chapter 9 discusses the ideas underlying the use of randomization and designed experiments, and closes with an account of mixed effect models, in which some parameters are treated as random. These two chapters are largely devoted to the classical linear model, in which the responses are supposed normally distributed, but since around 1970 regression modelling has greatly broadened. Chapter 10 is devoted to nonlinear models. It starts with an account of likelihood estimation using the iterative weighted least squares algorithm, which subsequently plays a unifying role, and then describes generalized linear models, binary data and loglinear models, semiparametric regression by local likelihood estimation and by penalized likelihood. It closes with an account of regression modelling of survival data.

Bayesian statistics is discussed in Chapter 11, starting with discussion of the role of prior information, followed by an account of Bayesian analogues of procedures developed in the earlier chapters. This is followed by a brief overview of Bayesian computation, including Laplace approximation, the Gibbs sampler and Metropolis–Hastings algorithm. The chapter closes with discussion of hierarchical and empirical Bayes and a very brief account of decision theory.

Likelihood is a favourite tool of statisticians but sometimes gives poor inferences. Chapter 12 describes some reasons for this, and outlines how conditional or marginal likelihoods can give better procedures.

The main links among the chapters of this book are shown in Figure 1.5.

Notation

The notation used in this book is fairly standard, but there are not enough letters in the Roman and Greek alphabets for total consistency. Greek letters generally denote parameters or other unknowns, with α largely reserved for error rates and confidence levels in connection with significance tests and

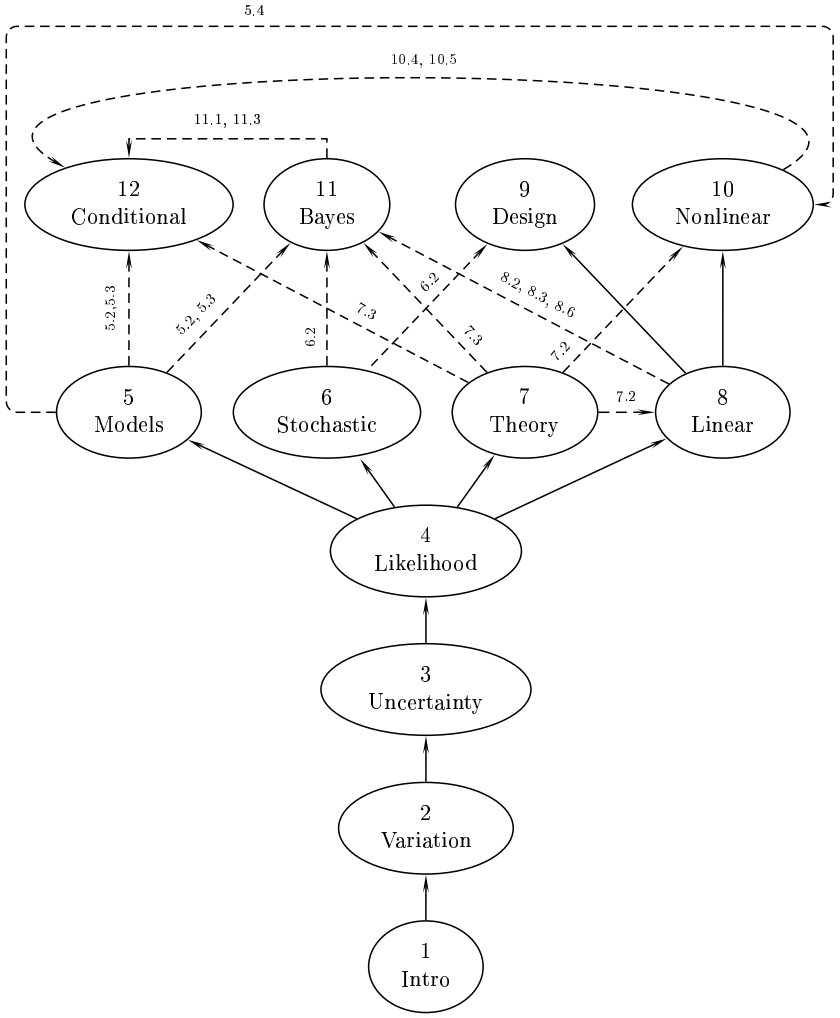

Figure 1.5 A map of the main dependencies among chapters of this book. A solid line indicates strong dependence and a dashed line indicates partial dependence through the given subsections.

confidence sets. Roman letters X, Y, Z, and so forth are mainly used for random variables, which take values x, y, z.

Probability, expectation, variance, covariance, and correlation are denoted $\Pr(\cdot)$, $\mathrm{E}(\cdot)$, $\mathrm{var}(\cdot)$ $\mathrm{cov}(\cdot,\cdot)$, and $\mathrm{corr}(\cdot,\cdot)$, while $\mathrm{cum}(\cdot,\cdot,\cdots)$ is occasionally used to denote a cumulant. We use $I(A)$ to denote the indicator random variable, which equals 1 if the event A occurs and 0 otherwise. A related function is the Heaviside function

$$H(u) = \begin{cases} 0, & u < 0, \\ 1, & u \geq 0, \end{cases}$$

whose generalized derivative is the Dirac delta function $\delta(u)$. This satisfies

$$\int \delta(y-u)g(u)\,du = g(y)$$

for any function g.

The Kronecker delta symbols δ_{rs}, δ_{rst}, and so forth all equal unity when all their subscripts coincide, and equal zero otherwise.

We use $\lfloor x \rfloor$ to denote the largest integer smaller than or equal to x, and $\lceil x \rceil$ to denote the smallest integer larger than or equal to x.

The symbol $\equiv$ indicates that constants have been dropped in defining a log likelihood, while $\doteq$ means 'approximately equals'. The symbols $\sim$, $\dot{\sim}$, $\overset{\text{ind}}{\sim}$, and $\overset{\text{iid}}{\sim}$ are shorthand for 'is distributed as', 'is approximately distributed as', 'are independently distributed as', and 'are independent and identically distributed as', while $\overset{D}{=}$ means 'has the same distribution as'. $X \perp Y$ means 'X is independent of Y'. We use $\xrightarrow{D}$ and $\xrightarrow{P}$ to denote convergence in distribution and in probability. To say that $Y_1, \ldots, Y_n$ are a random sample from some distribution means that they are independent and identically distributed according to that distribution.

We mostly reserve Z for standard normal random variables. As usual $N(\mu, \sigma^2)$ represents the normal distribution with mean μ and variance σ^2. The standard normal cumulative distribution and density functions are denoted Φ and ϕ. We use $c_\nu(\alpha)$, $t_\nu(\alpha)$, and $F_{\nu_1,\nu_2}(\alpha)$ to denote the α quantiles of the chi-squared distribution, Student t distribution with ν degrees of freedom, and F distribution with ν_1 and ν_2 degrees of freedom, while $U(0,1)$ denote the uniform distribution on the unit interval. Almost everywhere, z_α is the α quantile of the $N(0,1)$ distribution.

The data values in a sample of size n, typically denoted $y_1, \ldots, y_n$, are the observed values of the random variables $Y_1, \ldots, Y_n$; their average is $\overline{y} = n^{-1}\sum y_j$ and their sample variance is $s^2 = (n-1)^{-1}\sum (y_j - \overline{y})^2$.

We avoid boldface type, and rely on the context to make it plain when we are dealing with vectors or matrices; a^{T} denotes the matrix transpose of a vector or matrix a. The identity matrix of side n is denoted I_n, and 1_n is a $n \times 1$ vector of ones. If θ is a $p \times 1$ vector and $\ell(\theta)$ a scalar, then $\partial\ell(\theta)/\partial\theta$ is the $p \times 1$ vector whose rth element is $\partial\ell(\theta)/\partial\theta_r$, and $\partial^2\ell(\theta)/\partial\theta\partial\theta^{\mathrm{T}}$ is the $p \times p$ matrix whose (r,s) element is $\partial^2\ell(\theta)/\partial\theta_r\partial\theta_s$.

The end of each example is marked thus: ■

Exercise 2.1.3 denotes the third exercise at the end of Section 2.1, Problem 2.3 is the third problem at the end of Chapter 2, and so forth.

2

Variation

The key idea in statistical modelling is to treat the data as the outcome of a random experiment. The purpose of this chapter is to understand some consequences of this: how to summarize and display different aspects of random data, and how to use results of probability theory to appreciate the variation due to this randomness. We outline the elementary notions of statistics and parameters, and then describe how data and statistics derived from them vary under sampling from statistical models. Many quantities used in practice are based on averages or on ordered sample values, and these receive special attention. The final section reviews moments and cumulants, which will be useful in later chapters.

2.1 Statistics and Sampling Variation

2.1.1 Data summaries

The most basic element of data is a single observation, y — usually a number, but perhaps a letter, curve, or image. Throughout this book we shall assume that whatever their original form, the data can be recoded as numbers. We shall mostly suppose that single observations are scalar, though sometimes they are vectors or matrices.

We generally deal with an ensemble of n observations, $y_1, \ldots, y_n$, known as a *sample*. Occasionally interest centres on the given sample alone, and if n is not tiny it will be useful to summarize the data in terms of a few numbers. We say that a quantity $s = s(y_1, \ldots, y_n)$ that can be calculated from $y_1, \ldots, y_n$ is a *statistic*. Such quantities may be wanted for many different purposes.

Location and scale

Two basic features of a sample are its typical value and a measure of how spread out the sample is, sometimes known respectively as *location* and *scale*. They can be summarized in many ways.

Example 2.1 (Sample moments) Sample moments are calculated by putting mass n^{-1} on each of the y_j, and then calculating the mean, variance, and so forth. The simplest of these sample moments are

$$\overline{y} = \frac{1}{n}\sum_{j=1}^{n} y_j = \frac{1}{n}(y_1 + \cdots + y_n) \quad \text{and} \quad \frac{1}{n}\sum_{j=1}^{n}(y_j - \overline{y})^2;$$

we call the first of these the *average*. In practice the denominator n in the second moment is usually replaced by $n-1$, giving the *sample variance*

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_j - \overline{y})^2. \tag{2.1}$$

The denominator $n-1$ is justified in Example 2.14.

Here $\overline{y}$ and s have the same dimensions as the y_j, and are measures of location and scale respectively. ■

Potential confusion is avoided by using the word *average* to refer to a quantity calculated from data, and the words *mean* or *expectation* for the corresponding theoretical quantity; this convention is used throughout this book.

Example 2.2 (Order statistics) The *order statistics* of $y_1, \ldots, y_n$ are their values put in increasing order, which we denote $y_{(1)} \le y_{(2)} \le \cdots \le y_{(n)}$. If $y_1 = 5$, $y_2 = 2$ and $y_3 = 4$, then $y_{(1)} = 2$, $y_{(2)} = 4$ and $y_{(3)} = 5$. Examples of order statistics are the *sample minimum* $y_{(1)}$ and *sample maximum* $y_{(n)}$, and the lower and upper *quartiles* $y_{(\lceil n/4 \rceil)}$ and $y_{(\lceil 3n/4 \rceil)}$. The lowest quarter of the sample lies below the lower quartile, and the highest quarter lies above the upper quartile.

$\lceil u \rceil$ denotes the smallest integer greater than or equal to u.

Among statistics that can be based on the $y_{(j)}$ are the *sample median*, defined as

$$\text{median}(y_j) = \begin{cases} y_{((n+1)/2)}, & n \text{ odd}, \\ \frac{1}{2}\left(y_{(n/2)} + y_{(n/2+1)}\right), & n \text{ even}. \end{cases} \tag{2.2}$$

This is the centre of the sample: equal proportions of the data lie above and below it.

All these statistics are examples of *sample quantiles*. The pth sample quantile is the value with a proportion p of the sample to its left. Thus the minimum, maximum, quartiles, and median are (roughly) the 0, 1, 0.25, 0.75 and

0.5 sample quantiles. Like the median (2.2) when n is even, the pth sample quantile for non-integer pn is usually calculated by linear interpolation between the order statistics that bracket it.

Another measure of location is the average of the central observations of the sample. Suppose that p lies in the interval $[0, 0.5)$, and that $k = pn$ is an integer. Then the *$p\times 100\%$ trimmed average* is defined as

$$\frac{1}{n-2k}\sum_{j=k+1}^{n-k} y_{(j)},$$

which is the usual average $\overline{y}$ when $p = 0$. The 50% trimmed average ($p = 0.5$) is defined to be the median, while other values of p interpolate between the average and the median. Linear interpolation is used when pn is non-integer.

The statistics above measure different aspects of sample location. Some measures of scale based on the order statistics are the *range*, $y_{(n)} - y_{(1)}$, the *interquartile range* and the *median absolute deviation*,

$$\text{IQR} = y_{(\lceil 3n/4 \rceil)} - y_{(\lceil n/4 \rceil)}, \quad \text{MAD} = \text{median}\left\{|y_i - \text{median}(y_j)|\right\}.$$

These are, respectively, the difference between the largest and smallest observations, the difference between the observations at the ends of the central 50% of the sample, and the median of the absolute deviations of the observations from the sample median. One would expect the range of a sample to grow with its size, but the IQR and MAD should depend less on the sample size and in this sense are more stable measures of scale. ■

It is easy to establish that the mapping $y_1, \ldots, y_n \mapsto a + by_1, \ldots, a + by_n$ changes the values of location and scale measures in the previous examples by $m, s \mapsto a + bm, bs$ (Exercise 2.1.1); this seems entirely reasonable.

Bad data

The statistics described in Examples 2.1 and 2.2 measure different aspects of location and of scale. They also differ in their susceptibility to bad data. Consider what happens when an error, due perhaps to mistyping, results in an observation that is unusual compared to the others — an *outlier*. If the 'true' y_1 is replaced by $y_1 + \delta$, the average changes from $\overline{y}$ to $\overline{y} + n^{-1}\delta$, which could be arbitrarily large, while the sample median changes by a bounded amount — the most that can happen is that it moves to an adjacent observation. We say that the sample median is *resistant*, while the average is not. Roughly a quarter of the data would have to be contaminated before the interquartile range could change by an arbitrarily large amount, while the range and sample variance are sensitive to a single bad observation. The large-sample proportion of contaminated observations needed to change the value of a statistic by an arbitrarily large amount is called its *breakdown point*; it is a common measure of the resistance of a statistic.

Table 2.1 Seven successive days of times (hours) spent by women giving birth in the delivery suite at the John Radcliffe Hospital. (Data kindly supplied by Ethel Burns.)

Woman	Day						
	1	2	3	4	5	6	7
1	2.10	4.00	2.60	1.50	2.50	4.00	2.00
2	3.40	4.10	3.60	4.70	2.50	4.00	2.70
3	4.25	5.00	3.60	4.70	3.40	5.25	2.75
4	5.60	5.50	6.40	7.20	4.20	6.10	3.40
5	6.40	5.70	6.80	7.25	5.90	6.50	4.20
6	7.30	6.50	7.50	8.10	6.25	6.90	4.30
7	8.50	7.25	7.50	8.50	7.30	7.00	4.90
8	8.75	7.30	8.25	9.20	7.50	8.45	6.25
9	8.90	7.50	8.50	9.50	7.80	9.25	7.00
10	9.50	8.20	10.40	10.70	8.30	10.10	9.00
11	9.75	8.50	10.75	11.50	8.30	10.20	9.25
12	10.00	9.75	14.25		10.25	12.75	10.70
13	10.40	11.00	14.50		12.90	14.60	
14	10.40	11.20			14.30		
15	16.00	15.00					
16	19.00	16.50					

Example 2.3 (Birth data) Table 2.1 shows data extracted from a census of all the women who arrived to give birth at the John Radcliffe Hospital in Oxford during a three-month period. The table gives the times that women with vaginal deliveries —that is, without caesarian section — spent in the delivery suite, for the first seven of 92 successive days of data.

The initial step in dealing with data is to scrutinize them closely, and to understand how they were collected. In this case the time for each birth was recorded by the midwife who attended it, and numerous problems might have arisen in the recording. For example, one midwife might intend 4.20 to mean 4.2 hours, but another might mean 4 hours and 20 minutes. Moreover it is difficult to believe that a time can be known as exactly as 2 hours and 6 minutes, as would be implied by the value 2.10. Furthermore, there seems to be a fair degree of rounding of the data. In fact the data collection form was carefully prepared, and the midwives were trained in how to compile it, so the data are of high quality. Nevertheless it is important always to ask how the data were collected, and if possible to see the process at work.

Ideally the statistician assists in deciding what data are collected, and how.

The average of the $n = 95$ times in Table 2.1 is $\overline{y} = 7.57$ hours. The variance of the time spent in the delivery suite can be estimated by the sample variance, $s^2 = 12.97$ squared hours. The minimum, median, and maximum are 1.5, 7.5 and 19 hours respectively, and the quartiles are 4.95 and 9.75 hours. The 0.2 and 0.4 trimmed averages, 7.48 and 7.55 hours, are similar to $\overline{y}$ because there are no gross outliers. ■

Shape

The shape of a sample is also important. For example, the upper tails of annual income distributions are typically very fat, because a few individuals earn enormously more than most of us. The shape of such a distribution can be used to assess inequality, for example by considering the proportion of individuals whose annual income is less than one-half the median. Since shape does not depend on location or scale, statistics intended to summarize it should be invariant to location and scale shifts of the data.

Example 2.4 (Sample skewness) One measure of shape is the *standardized sample skewness*,

$$g_1 = \frac{n^{-1}\sum_{j=1}^{n}(y_j - \overline{y})^3}{\left\{(n-1)^{-1}\sum_{j=1}^{n}(y_j - \overline{y})^2\right\}^{3/2}}.$$

If the data are perfectly symmetric, $g_1 = 0$, while if they have a heavy upper tail, $g_1 > 0$, and conversely. For the times in the delivery suite, $g_1 = 0.65$: the data are somewhat skewed to the right. ■

Example 2.5 (Sample shape) Measures of shape can also be based on the sample quantiles. One is $(y_{(\lceil 0.95n \rceil)} - y_{(\lceil 0.5n \rceil)})/(y_{(\lceil 0.5n \rceil)} - y_{(\lceil 0.05n \rceil)})$, which takes value one for a symmetric distribution, and is more resistant to outliers than is the sample skewness. For the times in the delivery suite, this is 1.43, again showing skewness to the right. A value less than one would indicate skewness to the left. ■

It is straightforward to show that both these statistics are invariant to changes in the location and scale of $y_1, \ldots, y_n$.

Graphs

Graphs are indispensable in data analysis, because the human visual system is so good at recognizing patterns that the unexpected can leap out and hit the investigator between the eyes. An adverse effect of this ability is that patterns may be imagined even when they are absent, so experience, often aided by suitable statistics, is needed to interpret a graph. As any plot can be represented numerically, it too is a statistic, though to treat it merely as a set of numbers misses the point.

This can lead to inter-ocular trauma.

Example 2.6 (Histogram) Perhaps the best-known statistical graph is the *histogram*, constructed from scalar data by dividing the horizontal axis into disjoint bins — the intervals $I_1, \ldots, I_K$ — and then counting the observations in each. Let n_k denote the number of observations in I_k, for $k = 1, \ldots, K$, so $\sum_k n_k = n$. If the bins have equal width δ, then $I_k = [L + (k-1)\delta, L + k\delta)$, where L, δ, and K are chosen so that all the y_j lie between L and $L + K\delta$.

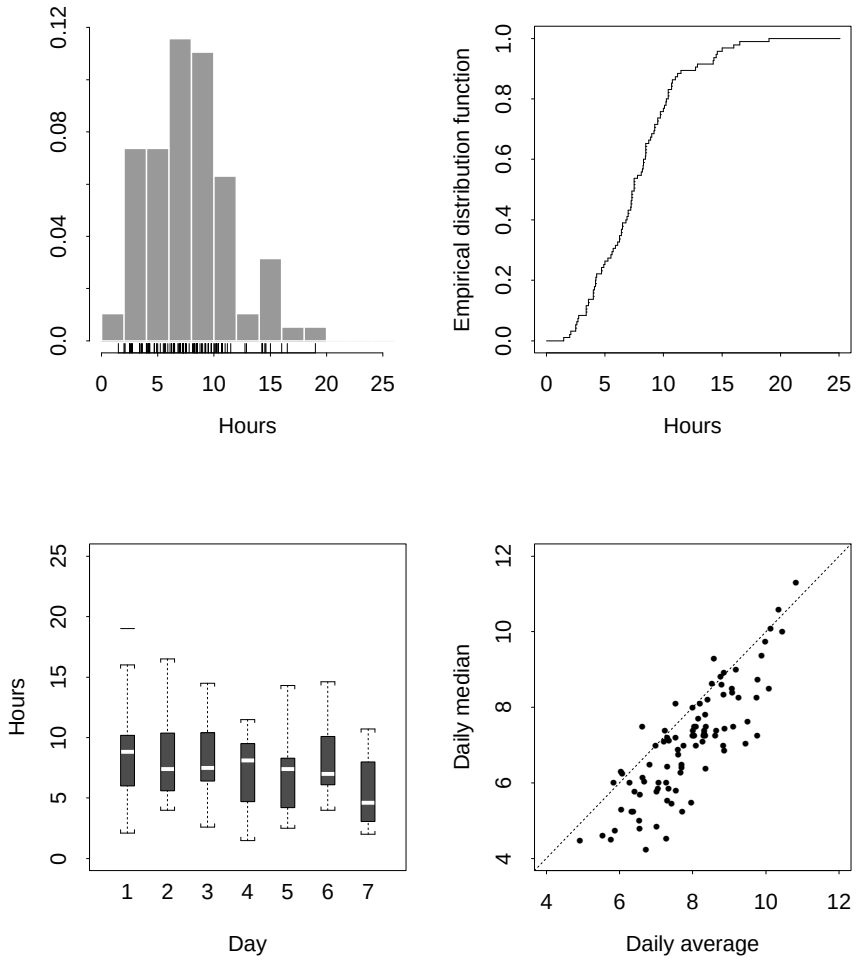

Figure 2.1
Summary plots for times in the delivery suite, in hours. Clockwise from top left: histogram, with rug showing values of observations; empirical distribution function; scatter plot of daily average hours against daily median hours, for all 92 days of data, with a line of unit slope through the origin; and boxplots for the first seven days.

We then plot the proportion n_k/n of the data in each bin as a column over it, giving the probability density function for a discretized version of the data.

The upper left panel of Figure 2.1 shows this for the birth data in Table 2.1, with $L = 0$, $\delta = 2$, and $K = 13$; the *rug* of tickmarks shows the data values themselves. As we would expect from Examples 2.4 and 2.5, the plot shows a density skewed to the right, with the most popular values in the range 5–10 hours. To increase δ would give fewer, wider, bins, while decreasing δ would give more, narrower, bins. It might be better to vary the bin width, with narrower bins in the centre of the data, and wider ones at the tails. ■

Example 2.7 (Empirical distribution function) The *empirical distribution function* (EDF) is the cumulative probability distribution that puts probability n^{-1} at each of $y_1, \ldots, y_n$. This is expressed mathematically as

$$n^{-1} \sum_{j=1}^{n} H(y - y_j), \tag{2.3}$$

where the distribution function that puts mass one at $u = 0$, that is,

$$H(u) = \begin{cases} 0, & u < 0, \\ 1, & u \geq 0, \end{cases}$$

is known as the Heaviside function. The EDF is a step function that jumps by n^{-1} at each of the y_j; of course it jumps by more at values that appear in the sample several times.

The upper right panel of Figure 2.1 shows the EDF of the times in the delivery suite. It is more detailed than the histogram, but perhaps conveys less information about the shape of the data. Which is preferable is partly a matter of taste, and depends on the use to which they will be put. ■

Example 2.8 (Scatterplot) When an observation has two components, $y_j = (u_j, v_j)$, a *scatter plot* is a plot of the v_j on the vertical axis against the u_j on the horizontal axis. An example is given in the lower right panel of Figure 2.1, which shows the median daily time in the delivery suite plotted against the average daily time, for the full 92 days for which data are available. As most points lie below the line with unit slope, and as the slope of the point cloud is slightly greater than one, the medians are generally smaller and somewhat more variable than the averages. The average and sample variance of the medians are 7.03 hours and 2.15 hours squared; the corresponding figures for the averages are 7.90 and 1.54. ■

Example 2.9 (Boxplot) Boxplots are usually used to compare related sets of data. An illustration is in the lower left panel of Figure 2.1, which compares the hours in the delivery suite for the seven different days in Table 2.1. For each day, the ends of the central box show the quartiles and the white line in its centre represents the daily median: thus about one-half of the data lie in the box, and its length shows the interquartile range IQR for that day. The bracket above the box shows the largest observation less than or equal to the upper quartile plus 1.5IQR. Likewise the bracket below shows the smallest observation greater than or equal to the lower quartile minus 1.5IQR. Values outside the brackets are plotted individually. The aim is to give a good idea of the location, scale, and shape of the data, and to show potential outliers clearly, in order to facilitate comparison of related samples. Here, for example, we see that the daily median varies from 5–10 hours, and that the daily IQR is fairly stable. ■

It takes thought to make good graphs. Some points to bear in mind are:

- the data should be made to stand out, in particular by avoiding so-called *chart-junk* — unnecessary labels, lines, shading, symbols and so forth;
- the axis labels and caption should make the graph as self-explanatory as possible, in particular containing the names and units of measurement of variables;
- comparison of related quantities should be made easy, for example by using identical scales of measurement, and placing plots side by side;
- scales should be chosen so that the most important systematic relations between variables are at about 45° to the axes;
- the *aspect ratio* — the ratio of the height of a plot to its width — can be varied to highlight different features of the data;
- graphs should be laid out so that departures from 'standard' appear as departures from linearity or from random scatter; and
- major differences in the precision of points should be indicated, at least roughly.

Perception experiments have shown that the eye is best at judging departures from 45°.

Nowadays it is easy to produce graphs, but unfortunately even easier to produce bad ones: there is no substitute for drafting and redrafting each graph to make it as clear and informative as possible.

2.1.2 Random sample

So far we have supposed that the sample $y_1, \ldots, y_n$ is of interest for its own sake. In practice, however, data are usually used to make inferences about the system from which they came. One reason for gathering the birth data, for example, was to assess how the delivery suite should be staffed, a task that involves predicting the patterns with which women will arrive to give birth, and how long they are likely to stay in the delivery suite once they are there. Though it is not useful to do this for births that have already occurred, the data available can help in making predictions, provided we can forge a link between the past and future. This is one use of a statistical model.

The fundamental idea of statistical modelling is to treat data as the observed values of random variables. The most basic model is that the data $y_1, \ldots, y_n$ available are the observed values of a *random sample of size* n, defined to be a collection of n independent identically distributed random variables, $Y_1, \ldots, Y_n$. We suppose that each of the Y_j has the same cumulative distribution function, F, which represents the population from which the sample has been taken. If F were known, we could in principle use the rules of probability calculus to deduce any of its properties — such as its mean and variance, or the probability distribution for a future observation — and any difficulties would be purely computational. In practice, however, F

Or sometimes a *simple random sample.*

is unknown, and we must try to infer its properties from the data. Often the quantity of central interest is a nonrandom function of F, such as its mean or its p quantile,

$$\mathrm{E}(Y) = \int y\, dF(y), \quad y_p = F^{-1}(p) = \inf\{y : F(y) \geq p\}; \tag{2.4}$$

these are the population analogues of the sample average and quantiles defined in Examples 2.1 and 2.2. Often there is a simple form for F^{-1} and the infimum is unnecessary. Other population quantities such as the interquartile range, $F^{-1}\left(\frac{3}{4}\right) - F^{-1}\left(\frac{1}{4}\right)$, are defined similarly.

We use $dF(y)$ to accommodate the possibility that F is discrete. If it bothers you, take $dF(y) = f(y)\, dy$.

Example 2.10 (Laplace distribution) A random variable Y for which

$$f(y; \eta, \tau) = \frac{1}{2\tau} \exp\left(-|y - \eta|/\tau\right), \quad -\infty < y < \infty,\ -\infty < \eta < \infty, \tau > 0, \tag{2.5}$$

is said to have the Laplace distribution. As $f(\eta + u; \eta, \tau) = f(\eta - u; \eta, \tau)$ for any u, the density is symmetric about η. Its integral is clearly finite, so $\mathrm{E}(Y) = \eta$, and evidently its median $y_{0.5} = \eta$ also. Its variance is

$$\mathrm{var}(Y) = \frac{1}{2\tau} \int_{-\infty}^{\infty} (y - \eta)^2 \exp\left(-|y - \eta|/\tau\right)\, dy = \tau^2 \int_0^{\infty} u^2 e^{-u}\, du = 2\tau^2,$$

as follows after the substitution $u = (y - \eta)/\tau$ and integration by parts; see Exercise 2.1.3. Integration of (2.5) gives

$$F(y) = \begin{cases} \frac{1}{2} \exp\left\{(y - \eta)/\tau\right\}, & y \leq \eta, \\ 1 - \frac{1}{2} \exp\left\{-(y - \eta)/\tau\right\}, & y > \eta, \end{cases}$$

so

$$F^{-1}(p) = \begin{cases} \eta + \tau \log(2p), & p < \frac{1}{2}, \\ \eta - \tau \log\{2(1 - p)\}, & p \geq \frac{1}{2}, \end{cases}$$

the interquartile range is

$$F^{-1}\left(\tfrac{3}{4}\right) - F^{-1}\left(\tfrac{1}{4}\right) = \eta + \tau \log 2 - (\eta - \tau \log 2) = 2\tau \log 2,$$

and the median absolute deviation is $\tau \log 2$ (Exercise 2.1.5). ■

Pierre-Simon Laplace (1749–1827) helped establish the metric system during the French Revolution but was dismissed by Napoleon 'because he brought the spirit of the infinitely small into the government' — presumably Bonaparte was unimpressed by differentiation. Laplace worked on celestial mechanics, published an important book on probability, and derived the least squares rule.

Quantities such as $\mathrm{E}(Y)$, $\mathrm{var}(Y)$ and $F^{-1}(p)$ are called *parameters*, and as their values depend on F, they are typically unknown. If F is determined by a finite number of parameters, θ, the model is *parametric*, and we may write $F = F(y; \theta)$, with corresponding probability density function $f(y; \theta)$. Ignorance about F then boils down to uncertainty about θ.

We use the term probability density function to mean the density function for a continuous variable, and the mass function for a discrete variable, and use the notation $f(y; \theta)$ in both cases.

It is natural to use sample quantities for inference about model parameters. Suppose that the data $Y_1, \ldots, Y_n$ are a random sample from a distribution F, that we are interested in a parameter θ that depends on F, and that we wish to use the statistic $S = s(Y_1, \ldots, Y_n)$ to make inferences about θ, for example

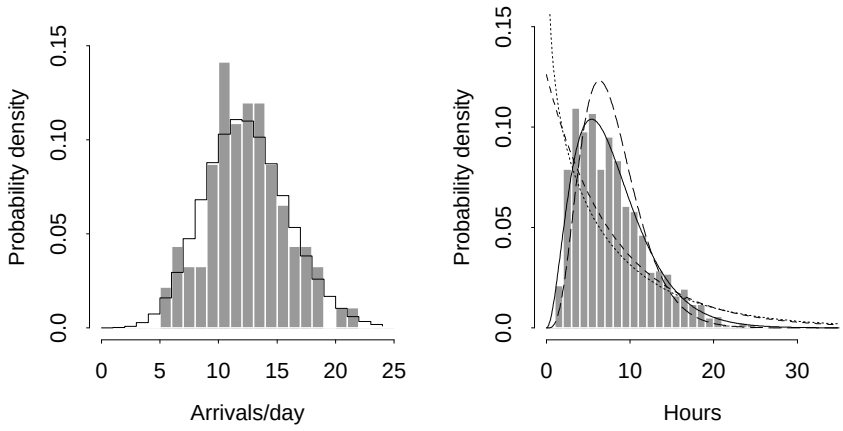

Figure 2.2
Comparisons of 92 days of delivery suite data with Poisson and gamma models. The left panel shows a histogram of the numbers of arrivals per day, with the PDF of the Poisson distribution with mean $\theta = 12.9$ overlaid. The right panel shows a histogram of the hours in the delivery suite for the 1187 births, with the PDFs of gamma distributions overlaid. The gamma distributions all have mean $\kappa/\lambda = 7.93$ hours. Their shape parameters are $\kappa = 3.15$ (solid), 0.8 (dots), 1 (small dashes), and 5 (large dashes).

hoping that S will be close to θ. Then we call S an *estimator* of θ and say that the particular value that S takes when the observed data are $y_1, \ldots, y_n$, that is, $s = s(y_1, \ldots, y_n)$, is an *estimate* of θ. This is the usual distinction between a random variable and the value that it takes, here S and s.

Siméon Denis Poisson (1781–1840) learned mathematics in Paris from Laplace and Lagrange. He did major work on definite integrals, on Fourier series, on elasticity and magnetism, and in 1837 published an important book on probability.

Example 2.11 (Poisson distribution) The Poisson distribution with mean θ has probability density function

$$\Pr(Y = y) = f(y;\theta) = \frac{\theta^y}{y!}e^{-\theta}, \quad y = 0, 1, 2, \ldots, \quad \theta > 0. \tag{2.6}$$

This discrete distribution is used for count data. For example, the left panel of Figure 2.2 shows a histogram of the number of women arriving at the delivery suite for each of the 92 days of data, together with the probability density function (2.6) with $\theta = 12.9$, equal to the average number of arrivals over the 92 days. This distribution seems to fit the data more or less adequately. ■

$\Gamma(\kappa)$ is the *gamma function*; see Exercise 2.1.3 for some of its properties.

Example 2.12 (Gamma distribution) The gamma distribution with scale parameter λ and shape parameter κ has probability density function

$$f(y;\lambda,\kappa) = \frac{\lambda^\kappa y^{\kappa-1}}{\Gamma(\kappa)}\exp(-\lambda y), \quad y > 0, \quad \lambda, \kappa > 0. \tag{2.7}$$

This distribution has mean κ/λ and variance κ/λ^2.

When $\kappa = 1$ the density is exponential, for $0 < \kappa < 1$ it is L-shaped, and for $\kappa > 1$ it falls smoothly on either side of its maximum. These shapes are illustrated in the right panel of Figure 2.2, which shows the hours in the delivery suite for the 1187 births that took place over the three months of data. In each case the mean of the density matches the data average of 7.93

hours; the value $\kappa = 3.15$ of the shape parameter was chosen to match the variance of the data by solving simultaneously the equations $\kappa/\lambda = 7.93$, $\kappa/\lambda^2 = 12.97$. Evidently the solid curve gives the best fit of those shown.

It is important to appreciate that the parametrization of F is not carved in stone. Here it might be better to rewrite (2.7) in terms of its mean $\mu = \kappa/\lambda$ and the shape parameter κ, in which case the density is expressed as

$$\frac{1}{\Gamma(\kappa)} \left(\frac{\kappa}{\mu}\right)^{\kappa} y^{\kappa-1} \exp(-\kappa y/\mu), \quad y > 0, \quad \mu, \kappa > 0, \tag{2.8}$$

with variance μ^2/κ. As functions of y the shapes of (2.7) and (2.8) are the same, but their expression in terms of parameters is not. The range of possible densities is the same for any 1–1 reparametrization of (κ, λ), so one might write the density in terms of two important quantiles, for example, if this made sense in the context of a particular application. The central issue in choice of parametrization is directness of interpretation in the situation at hand. ■

Example 2.13 (Laplace distribution) To express the Laplace density (2.5) in terms of its mean and variance η and $2\tau^2$, we set $\tau^2 = \sigma^2/2$, giving

$$\frac{1}{\sqrt{2}\sigma} \exp\left(-\sqrt{2}|y - \eta|/\sigma\right) \quad -\infty < y < \infty, \quad -\infty < \eta < \infty, \sigma > 0.$$

Its shape as a function of y is unchanged, but the new formula is uglier. ■

2.1.3 Sampling variation

If the data $y_1, \ldots, y_n$ are regarded as the observed values of random variables, then it follows that the sample and any statistics derived from it might have been different. However, although we would expect variation over possible sets of data, we would also expect to see systematic patterns induced by the underlying model. For instance, having inspected the lower left panel of Figure 2.1, we would be surprised to be told that the median hours in the delivery suite on day 8 was 15 hours, though any value between 5 and 10 hours would seem quite reasonable. From a statistical viewpoint, data have both a random and a systematic component, and one common goal of data analysis is to disentangle these as far as possible. In order to understand the systematic aspect, it makes sense to ask how we would expect a statistic $s(y_1, \ldots, y_n)$ to behave on average, that is, to try and understand the properties of the corresponding random variable, $S = s(Y_1, \ldots, Y_n)$.

Example 2.14 (Sample moments) Suppose that $Y_1, \ldots, Y_n$ is a random sample from a distribution with mean μ and variance σ^2. Then the average

$\overline{Y}$ has expectation and variance

$$\begin{aligned} \mathrm{E}(\overline{Y}) &= \mathrm{E}\left(\frac{1}{n}\sum_{j=1}^{n} Y_j\right) = \frac{n}{n}\mathrm{E}(Y_j) = \mu, \\ \mathrm{var}(\overline{Y}) &= \mathrm{var}\left(\frac{1}{n}\sum_{j=1}^{n} Y_j\right) = \frac{1}{n^2}\sum_{j=1}^{n}\mathrm{var}(Y_j) = \frac{\sigma^2}{n}, \end{aligned}$$

because the Y_j are independent identically distributed random variables. Thus the expected value of the random variable $\overline{Y}$ is the population mean μ.

To find the expectation of the sample variance $S^2 = (n-1)^{-1}\sum_j (Y_j - \overline{Y})^2$, note that

$$\begin{aligned} \sum_{j=1}^{n}(Y_j - \overline{Y})^2 &= \sum_{j=1}^{n}\left\{Y_j - \mu - (\overline{Y} - \mu)\right\}^2 \\ &= \sum_{j=1}^{n}(Y_j - \mu)^2 - 2\sum_{j=1}^{n}(Y_j - \mu)(\overline{Y} - \mu) + \sum_{j=1}^{n}(\overline{Y} - \mu)^2 \\ &= \sum_{j=1}^{n}(Y_j - \mu)^2 - 2n(\overline{Y} - \mu)^2 + n(\overline{Y} - \mu)^2 \\ &= \sum_{j=1}^{n}(Y_j - \mu)^2 - n(\overline{Y} - \mu)^2. \end{aligned}$$

As

$$\begin{aligned} \mathrm{E}\left\{(n-1)S^2\right\} &= n\mathrm{E}\left\{(Y_j - \mu)^2\right\} - n\mathrm{E}\left\{(\overline{Y} - \mu)^2\right\} \\ &= n\sigma^2 - n\sigma^2/n \\ &= (n-1)\sigma^2, \end{aligned}$$

we see that S^2 has expected value σ^2. This explains our use of the denominator $n-1$ when defining the sample variance s^2 in (2.1): the expectation of the corresponding random variable equals the population variance.

The birth data of Table 2.1 have $n = 95$, and the realized values of the random variables $\overline{Y}$ and S^2 are $\overline{y} = 7.57$ and $s^2 = 12.97$. Thus $\overline{y}$ has estimated variance $s^2/n = 12.97/95 = 0.137$ and estimated standard deviation $0.137^{1/2} = 0.37$. This suggests that the underlying 'true' mean μ of the population of times spent in the delivery suite by women giving birth is close to 7.6 hours. ■

Example 2.15 (Birth data) Figure 2.2 suggests the following simple model for the birth data. Each day the number N of women arriving to give birth is Poisson with mean θ. The jth of these women spends a time Y_j in the

delivery suite, where Y_j is a gamma random variable with mean μ and variance σ^2. The values of these parameters are $\theta \doteq 13$, $\mu \doteq 8$ hours and $\sigma^2 \doteq 13$ hours squared. The average time and median times spent, $\overline{Y} = N^{-1}\sum Y_j$ and M, vary from day to day, with the lower right panel of Figure 2.1 suggesting that $\mathrm{E}(M) < \mathrm{E}(\overline{Y})$ and $\mathrm{var}(M) > \mathrm{var}(\overline{Y})$, properties we shall see theoretically in Example 2.30. ■

Much of this book is implicitly or explicitly concerned with distinguishing random and systematic variation. The notions of sampling variation and of a random sample are central, and before continuing we describe a useful tool for comparison of data and a distribution.

2.1.4 Probability plots

It is often useful to be able to check graphically whether data $y_1, \ldots, y_n$ come from a particular distribution. Suppose that in addition to the data we had a random sample $x_1, \ldots, x_n$ known to be from F. In order to compare the shapes of the samples, we could sort them to get $y_{(1)} \leq \cdots \leq y_{(n)}$ and $x_{(1)} \leq \cdots \leq x_{(n)}$, and make a *quantile-quantile* or *Q-Q plot* of $y_{(1)}$ against $x_{(1)}$, $y_{(2)}$ against $x_{(2)}$, and so forth. A straight line would mean that $y_{(j)} = a + bx_{(j)}$, so that the shape of the samples was identical, while distinct curvature would indicate systematic differences between them. If the line was close to straight, we could be fairly confident that $y_1, \ldots, y_n$ looks like a sample from F — after all, it would have a shape similar to the sample $x_1, \ldots, x_n$ which is from F.

Quantile-quantile plots are helpful for comparison of two samples, but when comparing a single sample with a theoretical distribution it is preferable to use F directly in a *probability plot*, in which the $y_{(j)}$ are graphed against the *plotting positions* $F^{-1}\{j/(n+1)\}$. This use of the $j/(n+1)$ quantile of F is justified in Section 2.3 as an approximation to $\mathrm{E}(X_{(j)})$, where $X_{(j)}$ is the random variable of which $x_{(j)}$ is a particular value. For example, the jth plotting positions for the normal and exponential distributions $\Phi\{(x-\mu)/\sigma\}$ and $1 - e^{-\lambda x}$ are $\mu + \sigma\Phi^{-1}\{j/(n+1)\}$ and $-\lambda^{-1}\log\{1 - j/(n+1)\}$. When parameters such as μ, σ, and λ are unknown, the plotting positions used are for standardized distributions, here $\Phi^{-1}\{j/(n+1)\}$ and $-\log\{1 - j/(n+1)\}$, which are sometimes called *normal scores* and *exponential scores*. Probability plots for the normal distribution are particularly common in applications and are also called *normal scores plots*. The interpretation of a probability plot is aided by adding the straight line that corresponds to perfect fit of F.

Example 2.16 (Birth data) The top left panel of Figure 2.3 shows a probability plot to compare the 95 times in the delivery suite with the normal distribution. The distribution does not fit the largest and smallest observations, and the data show some upward curvature relative to the straight line.

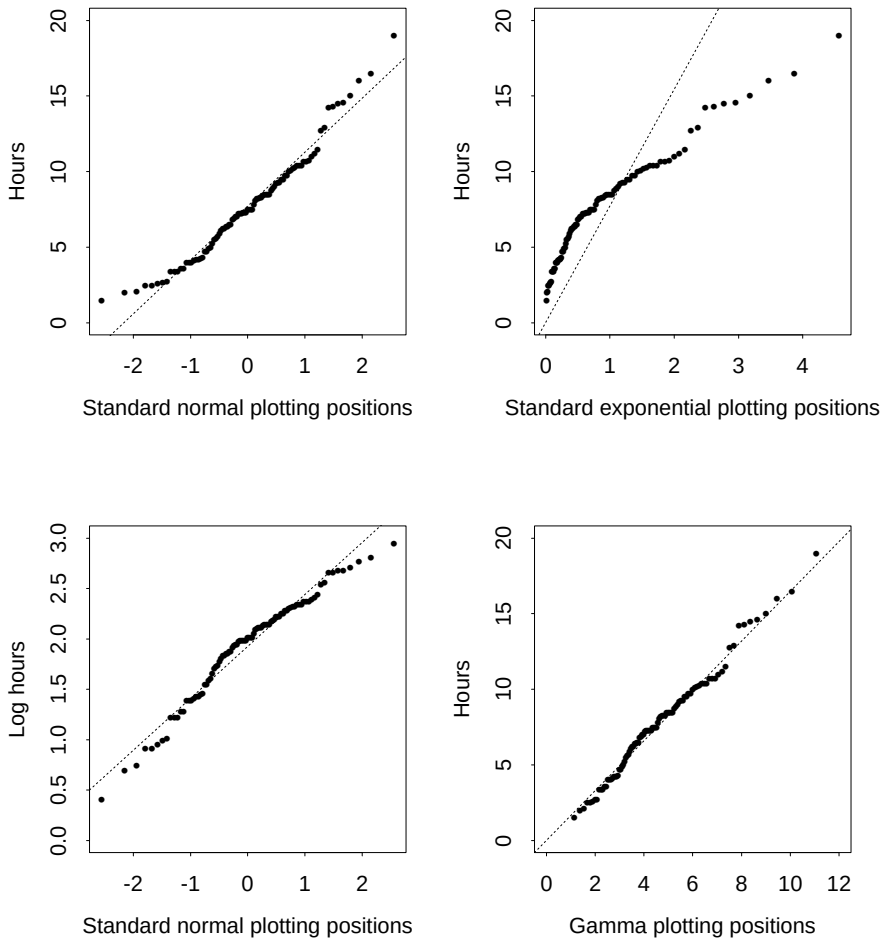

Figure 2.3
Probability plots for hours in the delivery suite, for the normal, exponential, gamma, and log-normal distributions (clockwise from top left). In each panel the dotted line is for a fitted distribution whose mean and variance match those of the data. None of the fits is perfect, but the gamma distribution fits best, and the exponential worst.

The top right panel shows that the exponential distribution would fit the data very poorly. The bottom left panel, a probability plot of the $\log y_j$ against normal plotting positions, corresponding to checking the log-normal distribution, shows slight downward curvature. The bottom right panel, a probability plot of the y_j against plotting positions for the gamma distribution with mean $\overline{y}$ and variance s^2, shows the best fit overall, though it is not perfect.

In the normal and gamma plots the dotted line corresponds to the theoretical distribution whose mean equals $\overline{y}$ and whose variance equals s^2; the dotted line in the exponential plot is for the exponential distribution whose mean equals $\overline{y}$; and the dotted line in the log-normal plot is for the normal

distribution whose mean and variance equal the average and variance of the $\log y_j$. ■

Some experience with interpreting probability plots may be gained from Practical 2.3.

Exercises 2.1

1 Let m and s be the values of location and scale statistics calculated from $y_1, \ldots, y_n$; m and s may be any of the quantities described in Examples 2.1 and 2.2. Show that the effect of the mapping $y_1, \ldots, y_n \mapsto a + by_1, \ldots, a + by_n$ $b > 0$, is to send $m, s \mapsto a + bm, bs$. Show also that the measures of shape in Examples 2.4 and 2.5 are unchanged by this transformation.

2 (a) Show that when δ is added to one of $y_1, \ldots, y_n$ and $|\delta| \to \infty$, the average $\overline{y}$ changes by an arbitrarily large amount, but the sample median does not. By considering such perturbations when n is large, deduce that the sample median has breakdown point 0.5.
(b) Find the breakdown points of the other statistics in Examples 2.1 and 2.2.

A sketch may help.

3 (a) If $\kappa > 0$ is real and k a positive integer, show that the gamma function

$$\Gamma(\kappa) = \int_0^\infty u^{\kappa-1} e^{-u}\, du,$$

has properties $\Gamma(1) = 1$, $\Gamma(\kappa + 1) = \kappa\Gamma(\kappa)$ and $\Gamma(k) = (k-1)!$. It is useful to know that $\Gamma(\frac{1}{2}) = \pi^{1/2}$, but you need not prove this.
(b) Use (a) to verify the mean and variance of (2.7).
(c) Show that for $0 < \kappa \leq 1$ the maximum value of (2.7) is at $y = 0$, and find its mode when $\kappa > 1$.

The *mode* of a density f is a value y such that $f(y) \geq f(x)$ for all x.

4 Give formulae analogous to (2.4) for the variance, skewness and 'shape' of a distribution F. Do they behave sensibly when a variable Y with distribution F is transformed to $a + bY$, so $F(y)$ is replaced by $F\{(y-a)/b\}$?

5 Let Y have continuous distribution function F. For any η, show that $X = |Y - \eta|$ has distribution $G(x) = F(\eta + x) - F(\eta - x)$, $x > 0$. Hence give a definition of the median absolute deviation of F in terms of F^{-1} and G^{-1}. If the density of Y is symmetric about the origin, show that $G(x) = 2F(x) - 1$. Hence find the median absolute deviation of the Laplace density (2.5).

6 A probability plot in which $y_1, \ldots, y_n$ and $x_1, \ldots, x_n$ are two random samples is called a *quantile-quantile* or *Q-Q plot*. Construct this plot for the first two columns in Table 2.1. Are the samples the same shape?

7 The *stem-and-leaf display* for the data 2.1, 2.3, 4.5, 3.3, 3.7, 1.2 is

```
1 | 2
2 | 13
3 | 37
4 | 5
```

If you turn the page on its side this gives a histogram showing the data values themselves (perhaps rounded); the units corresponding to intervals $[1, 2)$, $[2, 3)$ and so forth are to the left of the vertical bars, and the digits are to the right.

Construct this for the combined data for days 1–3 in Table 2.1. Hence find their median, quartiles, interquartile range, and range.

8 Do Figures 2.1–2.3 follow the advice given on page 23? If not, how could they be improved? Browse some textbooks and newspapers and think critically about any statistical graphics you find.

2.2 Convergence

2.2.1 Modes of convergence

Intuition tells us that the bigger our sample, the more faith we can have in our inferences, because our sample is more representative of the distribution F from which it came — if the sample size n was infinite, we would effectively know F. As $n \to \infty$ we can think of our sample $Y_1, \ldots, Y_n$ as converging to F, and of a statistic $S = s(Y_1, \ldots, Y_n)$ as converging to a limit that depends on F. For our purposes there are two main ways in which a sequence of random variables, $S_1, S_2, \ldots$, can converge to another random variable S.

Convergence in probability

We say that S_n *converges in probability* to S, $S_n \xrightarrow{P} S$, if for any $\varepsilon > 0$

$$\Pr(|S_n - S| > \varepsilon) \to 0 \quad \text{as} \quad n \to \infty. \tag{2.9}$$

A special case of this is the *weak law of large numbers*, whose simplest form is that if $Y_1, Y_2, \ldots$ is a sequence of independent identically distributed random variables each with finite mean μ, and if $\overline{Y} = n^{-1}(Y_1 + \cdots + Y_n)$ is the average of $Y_1, \ldots, Y_n$, then $\overline{Y} \xrightarrow{P} \mu$. We sometimes call this simply the *weak law*. It is illustrated in the left-hand panels of Figure 2.4, which show histograms of 10,000 averages of random samples of n exponential random variables, with $n = 1$, 5, 10, and 20. The individual variables have density e^{-y} for $y > 0$, so their mean μ and variance σ^2 both equal one. As n increases, the values of $S_n = \overline{Y}$ become increasingly concentrated around μ, so as the figure illustrates, $\Pr(|S_n - \mu| > \varepsilon)$ decreases for each positive ε.

Statistics that converge in probability have some useful properties. For example, if s_0 is a constant, and h is a function continuous at s_0, then if $S_n \xrightarrow{P} s_0$, it follows that $h(S_n) \xrightarrow{P} h(s_0)$ (Exercise 2.2.1).

An estimator S_n of a parameter θ is *consistent* if $S_n \xrightarrow{P} \theta$ as $n \to \infty$, whatever the value of θ. Consistency is desirable, but a consistent estimator that has poor properties for any realistic sample size will be useless in practice.

Example 2.17 (Binomial distribution) A binomial random variable $R = \sum_{j=1}^{m} I_j$ counts the numbers of ones in the random sample $I_1, \ldots, I_m$, each of which has a Bernoulli distribution,

Jacob Bernoulli (1654–1705) was a member of a mathematical family split by rivalry. His major work on probability, *Ars Conjectandi*, was published in 1713, but he also worked on many other areas of mathematics.

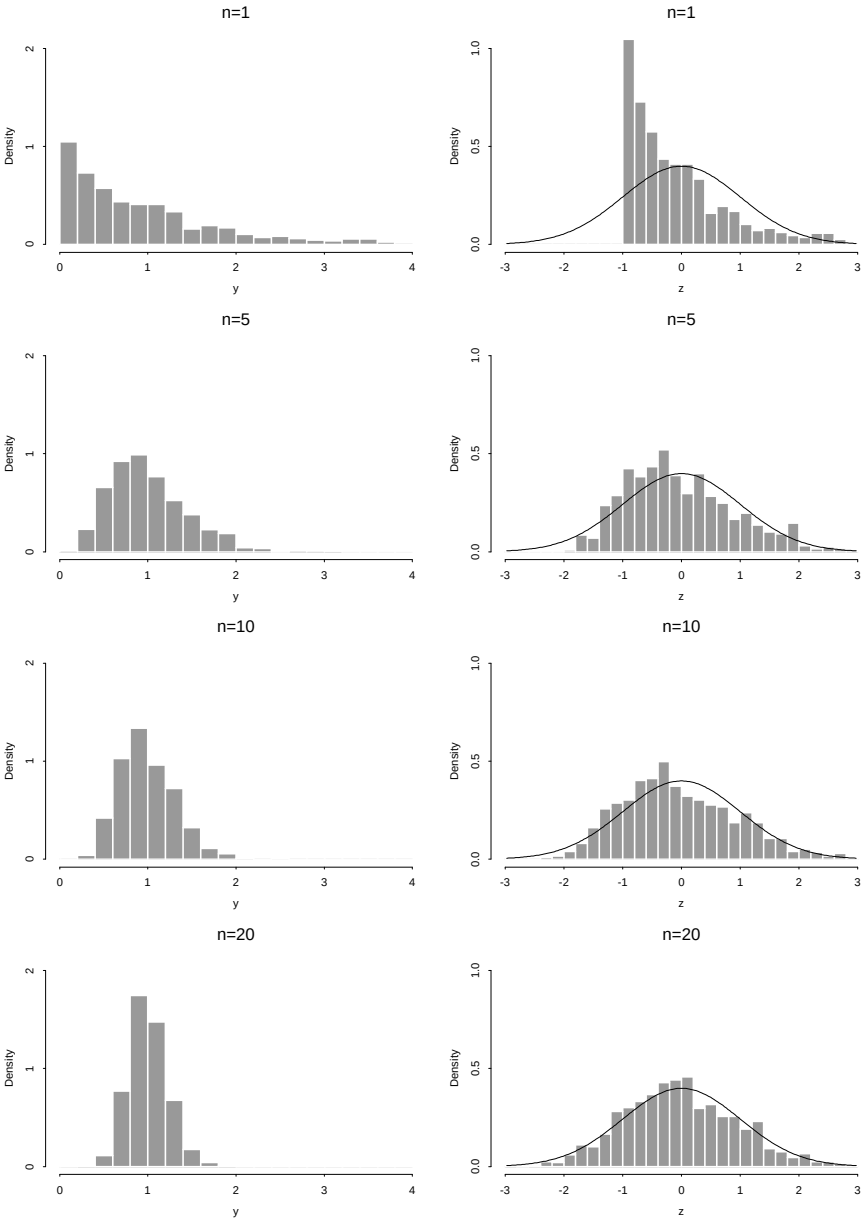

Figure 2.4
Convergence in probability and in distribution. The left panels show how histograms of the averages $\overline{Y}$ of 10,000 samples of n standard exponential random variables become more concentrated at the mean $\mu = 1$ as n increases through 1, 5, 10, and 20, due to the convergence in probability of $\overline{Y}$ to μ. The right panels show how the distribution of $Z_n = n^{1/2}(\overline{Y} - 1)$ approaches the standard normal distribution, due to the convergence in distribution of Z_n to normality.

$$\Pr(I_j = 1) = \pi,\ \Pr(I_j = 0) = 1 - \pi, \quad 0 \le \pi \le 1.$$

It is easy to check that $\mathrm{E}(I_j) = \pi$ and $\mathrm{var}(I_j) = \pi(1 - \pi)$. Thus the weak law applies to the proportion of successes $\widehat{\pi} = R/m$, giving $\widehat{\pi} \xrightarrow{P} \pi$ as $m \to \infty$. Evidently $\widehat{\pi}$ is a consistent estimator of π. However, the useless estimator

$\widehat{\pi} + 10^6/\log m$ is also consistent — consistency is a minimal requirement, not a guarantee that the estimator can safely be used in practice.

Each of the I_j has variance $\pi(1-\pi)$, and this is estimated by $\widehat{\pi}(1-\widehat{\pi})$, a continuous function of $\widehat{\pi}$ that converges in probability to $\pi(1-\pi)$. ■

Convergence in distribution

We say that the sequence $Z_1, Z_2, \ldots$, *converges in distribution* to Z, $Z_n \xrightarrow{D} Z$, if

$$\Pr(Z_n \leq z) \;\to\; \Pr(Z \leq z) \quad \text{as} \quad n \to \infty \tag{2.10}$$

at every z for which the distribution function $\Pr(Z \leq z)$ is continuous. The most important case of this is the *central limit theorem*, whose simplest version applies to a sequence of independent identically distributed random variables $Y_1, Y_2, \ldots$, with finite mean μ and finite variance $\sigma^2 > 0$. If the sample average is $\overline{Y} = n^{-1}(Y_1 + \cdots + Y_n)$, the Central Limit Theorem states that

$$Z_n = n^{1/2}\frac{(\overline{Y} - \mu)}{\sigma} \quad \xrightarrow{D} \quad Z, \tag{2.11}$$

where Z is a standard normal random variable, that is, one having the normal distribution with mean zero and variance one, written $N(0,1)$; see Section 3.2.1.

The right panels of Figure 2.4 illustrate such convergence. They show histograms of Z_n for the averages in the left-hand panels, with the standard normal probability density function superimposed. Each of the right-hand panels is a translation to zero of the histogram to its left, followed by 'zooming in': multiplication by a scale factor $n^{1/2}/\sigma$. As n increases, Z_n approaches its limiting standard normal distribution.

Example 2.18 (Average) Consider the average $\overline{Y}$ of a random sample with mean μ and finite variance $\sigma^2 > 0$. The weak law implies that $\overline{Y}$ is a consistent estimator of its expected value μ, and (2.11) implies that in addition $\overline{Y} = \mu + n^{-1/2}\sigma Z_n$, where $Z_n \xrightarrow{D} Z$. This supports our intuition that $\overline{Y}$ is a better estimate of μ for large n, and makes explicit the rate at which $\overline{Y}$ converges to μ: in large samples $\overline{Y}$ is essentially a normal variable with mean μ and variance σ^2/n. ■

Example 2.19 (Empirical distribution function) Let $Y_1, \ldots, Y_n$ be a random sample from F, and let $I_j(y)$ be the indicator random variable for the event $Y_j \leq y$. Thus $I_j(y)$ equals one if $Y_j \leq y$ and zero otherwise. The empirical distribution function of the sample is

$$\widehat{F}(y) = n^{-1}\sum_{j=1}^{n} I_j(y),$$

a step function that increases by n^{-1} at each observation, as in the upper right panel of Figure 2.1. We thought of (2.3) as a summary of the data $y_1, \ldots, y_n$; $\widehat{F}(y)$ is the corresponding random variable.

The $I_j(y)$ are independent and each has the Bernoulli distribution with probability $\Pr\{I_j(y) = 1\} = F(y)$. Therefore $\widehat{F}(y)$ is an average of independent identically distributed variables and has mean $F(y)$ and variance $F(y)\{1 - F(y)\}/n$. At a value y for which $0 < F(y) < 1$,

$$\widehat{F}(y) \xrightarrow{P} F(y), \quad n^{1/2}\frac{\{\widehat{F}(y) - F(y)\}}{[F(y)\{1 - F(y)\}]^{1/2}} \xrightarrow{D} Z, \text{ as } n \to \infty, \tag{2.12}$$

where Z is a standard normal variate. It can be shown that this pointwise convergence for each y extends to convergence of the function $\widehat{F}(y)$ to $F(y)$. The empirical distribution function in Figure 2.1 is thus an estimate of the true distribution of times in the delivery suite. ■

The alert reader will have noticed a sleight-of-word in the previous sentence. Convergence results tell us what happens as $n \to \infty$, but in practice the sample size is fixed and finite. How then are limiting results relevant? They are used to generate approximations for finite n — for example, (2.12) leads us to hope that $n^{1/2}\{\widehat{F}(y) - F(y)\}/[F(y)\{1 - F(y)\}]^{1/2}$ has approximately a standard normal distribution even when n is quite small. In practice it is important to check the adequacy of such approximations, and to develop a feel for their accuracy. This may be done analytically or by simulation (Section 3.3), while numerical examples are also valuable.

Slutsky's lemma

Convergence in distribution is useful in statistical applications because we generally want to compare probabilities. It is weaker than convergence in probability because it does not involve the joint distribution of S_n and S. If s_0 and u_0 are constants, these modes of convergence are related as follows:

$$S_n \xrightarrow{P} S \quad \Rightarrow \quad S_n \xrightarrow{D} S, \tag{2.13}$$

$$S_n \xrightarrow{D} s_0 \quad \Rightarrow \quad S_n \xrightarrow{P} s_0, \tag{2.14}$$

$$S_n \xrightarrow{D} S \text{ and } U_n \xrightarrow{P} u_0 \quad \Rightarrow \quad S_n + U_n \xrightarrow{D} S + u_0, S_nU_n \xrightarrow{D} Su_0. \tag{2.15}$$

The third of these is known as *Slutsky's lemma*.

Evgeny Evgenievich Slutsky (1880–1948) made fundamental contributions to stochastic convergence and to economic time series during the 1920s and 1930s. In 1902 he was expelled from university in Kiev for political activity. He studied in Munich and Kiev and worked in Kiev and Moscow.

Devotees of tricky analysis will find references to proofs of (2.13)–(2.15) in Section 2.5.

Example 2.20 (Sample variance) Suppose that $Y_1, \ldots, Y_n$ is a random sample of variables with finite mean μ and variance σ^2. Let

$$S_n = n^{-1}\sum_{j=1}^{n}(Y_j - \overline{Y})^2 = n^{-1}\sum_{j=1}^{n} Y_j^2 - \overline{Y}^2,$$

where $\overline{Y}$ is the sample average. The weak law implies that $\overline{Y} \xrightarrow{P} \mu$, and the function $h(x) = x^2$ is continuous everywhere, so $\overline{Y}^2 \xrightarrow{P} \mu^2$. Moreover

$$\mathrm{E}(Y_j^2) = \mathrm{var}(Y_j) + \{\mathrm{E}(Y_j)\}^2 = \sigma^2 + \mu^2,$$

so $n^{-1}\sum Y_j^2 \xrightarrow{P} \sigma^2+\mu^2$ also. Now (2.13) implies that $n^{-1}\sum Y_j^2 \xrightarrow{D} \sigma^2+\mu^2$, and therefore (2.15) implies that $S_n \xrightarrow{D} \sigma^2$. But σ^2 is constant, so $S_n \xrightarrow{P} \sigma^2$.

The sample variance S^2 may be written as $S_n \times n/(n-1)$, which evidently also tends in probability to σ^2. Thus not only is it true that for all n, $\mathrm{E}(S^2) = \sigma^2$, but the distribution of S^2 is increasingly concentrated at σ^2 in large samples. ■

These ideas extend to functions of several random variables.

Example 2.21 (Covariance and correlation) The covariance between random variables X and Y is

$$\gamma = \mathrm{E}[\{X - \mathrm{E}(X)\}\{Y - \mathrm{E}(Y)\}] = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y).$$

An estimate of γ based on a random sample of data pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ is the sample covariance

$$C = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X})(Y_j - \overline{Y}) = \frac{n}{n-1}\left(n^{-1}\sum_{j=1}^{n} X_jY_j - \overline{XY}\right),$$

where $\overline{X}$ and $\overline{Y}$ are the averages of the X_j and Y_j. Provided the moments $\mathrm{E}(XY)$, $\mathrm{E}(X)$ and $\mathrm{E}(Y)$ are finite, the weak law applies to each of $n^{-1}\sum X_jY_j$, $\overline{X}$ and $\overline{Y}$, which converge in probability to their expectations. The convergence is also in distribution, by (2.13), so (2.15) implies that $C \xrightarrow{D} \gamma$. But γ is constant, so (2.14) implies that $C \xrightarrow{P} \gamma$.

The correlation between X and Y,

$$\rho = \frac{\mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)}{\{\mathrm{var}(X)\mathrm{var}(Y)\}^{1/2}},$$

is such that $-1 \le \rho \le 1$. When $|\rho| = 1$ there is a linear relation between X and Y, so that $a + bX + cY = 0$ for some nonzero b and c (Exercise 2.2.3). Values of ρ close to ± 1 indicate strong linear dependence between the distributions of X and Y, though values close to zero do not indicate independence, just lack of a linear relation. The parameter ρ can be estimated from the pairs (X_j, Y_j) by the sample *correlation coefficient*,

Also known as the *product moment correlation coefficient.*

$$R = \frac{\sum_{j=1}^{n}(X_j - \overline{X})(Y_j - \overline{Y})}{\left\{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{k=1}^{n}(Y_k - \overline{Y})^2\right\}^{1/2}}.$$

The keen reader will enjoy showing that $R \xrightarrow{P} \rho$. ■

Example 2.22 (Studentized statistic) Suppose that $(T_n-\theta)/\mathrm{var}(T_n)^{1/2}$ converges in distribution to a standard normal random variable, Z, and that $\mathrm{var}(T_n) = \tau^2/n$, where $\tau^2 > 0$ is unknown but finite. Let V_n be a statistic that estimates τ^2/n, with the property that $nV_n \xrightarrow{P} \tau^2$. The function $h(x) = \tau/(nx)^{1/2}$ is continuous at $x = 1$, so $\tau/(nV_n)^{1/2} \xrightarrow{P} 1$. Therefore

$$Z_n = n^{1/2}\frac{(T_n-\theta)}{\tau} \times \frac{\tau}{(nV_n)^{1/2}} \xrightarrow{D} Z \times 1,$$

by (2.15). Thus Z_n has a limiting standard normal distribution provided that nV_n is a consistent estimator of τ^2.

The best-known instance of this is the average of a random sample, $\overline{Y} = n^{-1}(Y_1 + \cdots + Y_n)$. If the Y_j have finite mean θ and finite positive variance, σ^2, $\overline{Y}$ has mean θ and variance σ^2/n. The Central Limit Theorem states that

$$n^{1/2}\frac{(\overline{Y}-\theta)}{\sigma} \xrightarrow{D} Z.$$

Consider $Z_n = n^{1/2}(\overline{Y}-\theta)/S$, where $S^2 = (n-1)^{-1}\sum(Y_j-\overline{Y})^2$. Example 2.20 shows that $S^2 \xrightarrow{P} \sigma^2$, and it follows that $Z_n \xrightarrow{D} Z$.

The replacement of $\mathrm{var}(T_n)$ by an estimate is called studentization to honour W. S. Gossett. Publishing under the pseudonym 'Student' in 1908, he considered the effect of replacing σ by S for normal data; see Section 3.2. ■

William Sealy Gossett (1876–1937) worked at the Guinness brewery in Dublin. Apart from his contributions to beer and statistics, he also invented a boat with two rudders that would be easy to manoeuvre when fly fishing.

Intuition suggests that bigger samples always give better estimates, but intuition can mislead or fail.

Example 2.23 (Cauchy distribution) A Cauchy random variable centred at θ has density

$$f(y;\theta) = \frac{1}{\pi\{1+(y-\theta)^2\}}, \quad -\infty < y < \infty, \quad -\infty < \theta < \infty. \tag{2.16}$$

Although (2.16) is symmetric with mode at θ, none of its moments exist, and in fact the average $\overline{Y}$ of a random sample $Y_1, \ldots, Y_n$ of such data has the same distribution as a single observation. So if we were unlucky enough to have such a sample, it would be useless to estimate θ by $\overline{Y}$: we might as well use Y_1. The difficulty is that the tails of the Cauchy density decrease very slowly. Data with similar characteristics arise in many financial and insurance contexts, so this is not a purely mathematical issue: the average may be a poor estimate, and better ones are discussed later. ■

Augustin Louis Cauchy (1789–1857) made contributions to all the areas of mathematics known at his time. He was a pioneer of real and complex analysis, but also developed applied techniques such as Fourier transforms and the diagonalization of matrices in order to work on elasticity and the theory of light. His relations with contemporaries were often poor because of his rigid Catholicism and his difficult character.

2.2.2 Delta method

Variances and variance estimates are often required for smooth functions of random variables. Suppose that the quantity of interest is $h(T_n)$, and

$$(T_n - \mu)/\mathrm{var}(T_n)^{1/2} \xrightarrow{D} Z, \quad n\mathrm{var}(T_n) \xrightarrow{P} \tau^2 > 0,$$

as $n \to \infty$, and Z has the standard normal distribution. Then we may write $T_n = \mu + n^{-1/2}\tau Z_n$, where $Z_n \xrightarrow{D} Z$. If h has a continuous non-zero derivative h' at μ, Taylor series expansion gives

$$h(T_n) = h(\mu + n^{-1/2}\tau Z_n) = h(\mu) + n^{-1/2}\tau Z_n h'(\mu + n^{-1/2}\tau W_n),$$

where W_n lies between Z_n and zero. As h' is continuous at μ, it follows that $h'(\mu + n^{-1/2}\tau W_n) \xrightarrow{P} h'(\mu)$, so (2.15) gives

$$\begin{aligned} \frac{n^{1/2}\{h(T_n) - h(\mu)\}}{\tau h'(\mu)} &= \frac{n^{1/2}\{h(T_n) - h(\mu)\}}{\tau h'(\mu + n^{-1/2}\tau W_n)} \times \frac{h'(\mu + n^{-1/2}\tau W_n)}{h'(\mu)} \\ &= Z_n \times \frac{h'(\mu + n^{-1/2}\tau W_n)}{h'(\mu)} \\ &\xrightarrow{D} Z \end{aligned}$$

as $n \to \infty$. This implies that in large samples, $h(T_n)$ has approximately the normal distribution with mean $h(\mu)$ and variance $\mathrm{var}(T_n)h'(\mu)^2$, i.e.

$\dot{\sim}$ means 'is approximately distributed as'.

$$h(T_n) \;\dot{\sim}\; N(h(\mu), \mathrm{var}(T_n)h'(\mu)^2). \tag{2.17}$$

This result is often called the *delta method*. Analogous results apply if the limiting distribution of Z_n is non-normal.

Furthermore, if $h'(\mu)$ is replaced by $h'(T_n)$ and τ^2 is replaced by a consistent estimator, S_n, a modification of the argument in Example 2.22 gives

$$\frac{n^{1/2}\{h(T_n) - h(\mu)\}}{S_n^{1/2}|h'(T_n)|} \xrightarrow{D} Z. \tag{2.18}$$

Thus the same limiting results apply if the variance of $h(T_n)$ is replaced by a consistent estimator. In particular, replacement of the parameters in $\mathrm{var}(T_n)h'(\mu)^2$ by consistent estimators gives a consistent estimator of $\mathrm{var}\{h(T_n)\}$.

Example 2.24 (Exponential transformation) Consider $h(\overline{Y}) = \exp(\overline{Y})$, where $\overline{Y}$ is the average of a random sample of size n, and each of the Y_j has mean μ and variance σ^2. Here $h'(\mu) = e^\mu$, so $\exp(\overline{Y})$ is asymptotically normal with mean e^μ and variance $n^{-1}\sigma^2 e^{2\mu}$. This can be estimated by $n^{-1}S^2\exp(2\overline{Y})$, where S^2 is the sample variance. ■

Several variables

The delta method extends to functions of several random variables $T_1, \ldots, T_p$; we suppress dependence on n for ease of notation. As $n \to \infty$, suppose that for each r, $n^{-1/2}(T_r - \theta_r) \xrightarrow{D} N(0, \omega_{rr})$, that the joint limiting distribution of $n^{-1/2}(T_r - \theta_r)$ is multivariate normal (see Section 3.2.3) and $n\mathrm{cov}(T_r, T_s) \to \omega_{rs}$, where the $p \times p$ matrix Ω whose (r, s) element is ω_{rs} is positive-definite; note that Ω is symmetric. Now suppose that a variance is required for the scalar function $h(T_1, \ldots, T_p)$. An argument like that above gives

$$h(T_1, \ldots, T_p) \;\dot\sim\; N\left\{h(\theta_1, \ldots, \theta_p), n^{-1} h'(\theta)^{\mathrm{T}} \Omega h'(\theta)\right\}, \tag{2.19}$$

where $h'(\theta)$ is the $p \times 1$ vector whose rth element is $\partial h(\theta_1, \ldots, \theta_p)/\partial\theta_r$; the requirement that $h'(\theta) \neq 0$ also holds here. As in the univariate case, the variance can be estimated by replacing parameters with consistent estimators.

Example 2.25 (Ratio) Let $\theta_1 = \mathrm{E}(X) \neq 0$ and $\theta_2 = \mathrm{E}(Y)$, and suppose we are interested in $h(\theta_1, \theta_2) = \theta_2/\theta_1$. Estimates of θ_1 and θ_2 based on random samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ are $T_1 = \overline{X}$ and $T_2 = \overline{Y}$, so the ratio is consistently estimated by T_2/T_1. The derivative vector is $h'(\theta) = (-\theta_2/\theta_1^2, \theta_1^{-1})^{\mathrm{T}}$, and the limiting mean and variance of T_2/T_1 are

$$\frac{\theta_2}{\theta_1}, \quad n^{-1} \begin{pmatrix} -\theta_2/\theta_1^2 & \theta_1^{-1} \end{pmatrix} \begin{pmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22} \end{pmatrix} \begin{pmatrix} -\theta_2/\theta_1^2 \\ \theta_1^{-1} \end{pmatrix},$$

the second of which equals

$$(n\theta_1^2)^{-1} \left\{ \omega_{11} \left(\frac{\theta_2}{\theta_1}\right)^2 - 2\omega_{12}\frac{\theta_2}{\theta_1} + \omega_{22} \right\},$$

assumed finite and positive. The variance tends to zero as $n \to \infty$, so we should aim to estimate $n\mathrm{var}(T_2/T_1)$, which is not a moving target.

Examples 2.20 and 2.21 imply that ω_{11}, ω_{22}, and ω_{12} are consistently estimated by $S_1^2 = (n-1)^{-1}\sum(X_j - \overline{X})^2$, $S_2^2 = (n-1)^{-1}\sum(Y_j - \overline{Y})^2$, and $C = (n-1)^{-1}\sum(X_j - \overline{X})(Y_j - \overline{Y})$ respectively. Therefore $n\mathrm{var}(\overline{Y}/\overline{X})$ is consistently estimated by

$$\overline{X}^{-2} \left\{ S_1^2 \left(\frac{\overline{Y}}{\overline{X}}\right)^2 - 2C\frac{\overline{Y}}{\overline{X}} + S_2^2 \right\} = \frac{1}{(n-1)\overline{X}^2} \sum_{j=1}^{n} \left(Y_j - \frac{\overline{Y}}{\overline{X}} X_j \right)^2,$$

as we see after simplification. ■

Example 2.26 (Gamma shape) In Example 2.12 the shape parameter κ of the gamma distribution was taken to be $\overline{y}^2/s^2 = 3.15$, based on $n = 95$ observations. The corresponding random variable is T_1^2/T_2, where $T_1 = \overline{Y}$ and $T_2 = S^2$ are calculated from the random sample $Y_1, \ldots, Y_n$, supposed to be gamma with mean $\theta_1 = \kappa/\lambda$ and variance $\theta_2 = \kappa/\lambda^2$. We take $h(\theta_1, \theta_2) =$

θ_1^2/θ_2, giving $h'(\theta_1,\theta_2) = (2\theta_1/\theta_2, -\theta_1^2/\theta_2^2)^{\mathrm{T}}$. The variance of T_1 is θ_2/n, that is, $n^{-1}\kappa/\lambda^2$, and it turns out that

$$\mathrm{var}(T_2) = \mathrm{var}(S^2) = \frac{\kappa_4}{n} + \frac{2\kappa_2^2}{n-1}, \quad \mathrm{cov}(T_1, T_2) = \mathrm{cov}(\overline{Y}, S^2) = \frac{\kappa_3}{n},$$

where $\kappa_2 = \kappa/\lambda^2$, $\kappa_3 = 2\kappa/\lambda^3$, and $\kappa_4 = 6\kappa/\lambda^4$. Thus

$$\begin{aligned} \mathrm{var}(T_1^2/T_2) &\doteq \begin{pmatrix} 2\lambda & -\lambda^2 \end{pmatrix} \begin{pmatrix} \frac{\kappa}{n\lambda^2} & \frac{2\kappa}{n\lambda^3} \\ \frac{2\kappa}{n\lambda^3} & \frac{6\kappa}{n\lambda^4} + \frac{2\kappa^2}{(n-1)\lambda^4} \end{pmatrix} \begin{pmatrix} 2\lambda \\ -\lambda^2 \end{pmatrix} \\ &= \frac{2\kappa}{n}\left(1 + \frac{n\kappa}{n-1}\right), \end{aligned}$$

or roughly $2n^{-1}\kappa(\kappa+1)$. ■

This can be skipped on a first reading.

Big and little oh notation: O and o

For two sequences of constants, $\{s_n\}$ and $\{a_n\}$ such that $a_n \geq 0$ for all n, we write $s_n = o(a_n)$ if $\lim_{n\to\infty}(s_n/a_n) = 0$, and $s_n = O(a_n)$ if there is a finite constant k such that $\lim_{n\to\infty}|s_n| \leq a_n k$. A sequence of random variables $\{S_n\}$ is said to be $o_p(a_n)$ if $(S_n/a_n) \xrightarrow{P} 0$ as $n \to \infty$, and is said to be $O_p(a_n)$ if S_n/a_n is bounded in probability as $n \to \infty$, that is, given $\varepsilon > 0$ there exist n_0 and a finite k such that for all $n > n_0$,

$$\Pr(|S_n/a_n| < k) > 1 - \varepsilon.$$

This gives a useful shorthand for expansions of random quantities.

To illustrate this, suppose that $\{Y_j\}$ is a sequence of independent identically distributed variables with finite mean μ, and let $S_n = n^{-1}(Y_1 + \cdots + Y_n)$. Then the weak law may be restated as $S_n = \mu + o_p(1)$, and if in addition the Y_j have finite variance σ^2, the Central Limit Theorem implies that $\overline{Y} = \mu + O_p(n^{-1/2})$. More precisely, $\overline{Y} \stackrel{D}{=} \mu + n^{-1/2}\sigma Z + o_p(n^{-1/2})$, where Z has a standard normal distribution. Such expressions are sometimes used in later chapters.

$\stackrel{D}{=}$ means 'has the same distribution as'.

Exercises 2.2

1 Suppose that $S_n \xrightarrow{P} s_0$, and that the function h is continuous at s_0, that is, for any $\varepsilon > 0$ there exists a $\delta > 0$ such that $|x-y| < \delta$ implies that $|h(x)-h(y)| < \varepsilon$. Explain why this implies that

$$\Pr(|S_n - s_0| < \delta) \leq \Pr\{|h(S_n) - h(s_0)| < \varepsilon\} \leq 1,$$

and deduce that $\Pr\{|h(s_0)-h(S_n)| < \varepsilon\} \to 1$ as $n \to \infty$. That is, $h(S_n) \xrightarrow{P} h(s_0)$.

2 Let s_0 be a constant. By writing

$$\Pr(|S_n - s_0| \leq \varepsilon) = \Pr(S_n \leq s_0 + \varepsilon) - \Pr(S_n \leq s_0 - \varepsilon),$$

for $\varepsilon > 0$, show that $S_n \xrightarrow{D} s_0$ implies that $S_n \xrightarrow{P} s_0$.

3 (a) Let X and Y be two random variables with finite positive variances. Use the fact that $\mathrm{var}(aX+Y) \geq 0$, with equality if and only if the linear combination $aX+Y$ is constant with probability one, to show that $\mathrm{cov}(X,Y)^2 \leq \mathrm{var}(X)\mathrm{var}(Y)$; this is a version of the *Cauchy–Schwarz inequality*. Hence show that $-1 \leq \mathrm{corr}(X,Y) \leq 1$, and say under what conditions equality is attained.
(b) Show that if X and Y are independent, $\mathrm{corr}(X,Y) = 0$. Show that the converse is false by considering the variables X and $Y = X^2 - 1$, where X has mean zero, variance one, and $\mathrm{E}(X^3) = 0$.

4 Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ be independent random samples from the exponential densities $\lambda e^{-\lambda x}$, $x > 0$, and $\lambda^{-1} e^{-y/\lambda}$, $y > 0$, with $\lambda > 0$. If $\overline{X}$ and $\overline{Y}$ are the sample averages, show that $\overline{X}\,\overline{Y} \xrightarrow{P} 1$ as $n \to \infty$.

5 Show that as $n \to \infty$ the skewness measure in Example 2.4 converges in probability to the corresponding theoretical quantity

$$\frac{\int (y-\mu)^3 dF(y)}{\left\{\int (y-\mu)^2 dF(y)\right\}^{3/2}},$$

provided this has finite numerator and positive denominator. Under what additional condition(s) is the skewness measure asymptotically normal?

6 If $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, show that $n^{1/2}(\overline{Y} - \mu)^2 \xrightarrow{P} 0$ as $n \to \infty$. Given that $\mathrm{var}\{(Y_j - \mu)^2\} = 2\sigma^4$, deduce that $(S^2 - \sigma^2)/(2\sigma^4/n)^{1/2} \xrightarrow{D} Z$, where $Z \sim N(0,1)$. When is this true for non-normal data?

$\overset{\text{iid}}{\sim}$ means 'are independent and identically distributed as'.

7 Let R be a binomial variable with probability π and denominator m; its mean and variance are $m\pi$ and $m\pi(1-\pi)$. The *empirical logistic transform* of R is

$$h(R) = \log\left(\frac{R + \frac{1}{2}}{m - R + \frac{1}{2}}\right).$$

Show that for large m,

$$h(R) \sim N\left\{\log\left(\frac{\pi}{1-\pi}\right), \frac{1}{m\pi(1-\pi)}\right\}.$$

What is the exact value of $\mathrm{E}[\log\{R/(m-R)\}]$? Are the $\frac{1}{2}$s necessary in practice?

8 Truncated Poisson variables Y arise when counting quantities such as the sizes of groups, each of which must contain at least one element. The density is

$$\Pr(Y = y) = \frac{\theta^y e^{-\theta}}{y!\,(1 - e^{-\theta})}, \quad y = 1, 2, \ldots, \quad \theta > 0.$$

Find an expression for $\mathrm{E}(Y) = \mu(\theta)$ in terms of θ. If $Y_1, \ldots, Y_n$ is a random sample from this density and $n \to \infty$, show that $\overline{Y} \xrightarrow{P} \mu(\theta)$. Hence show that $\widehat{\theta} = \mu^{-1}(\overline{Y}) \xrightarrow{P} \theta$.

9 Let $Y = \exp(X)$, where $X \sim N(\mu, \sigma^2)$; Y has the log-normal distribution. Use the moment-generating function of X to show that $\mathrm{E}(Y^r) = \exp(r\mu + r^2\sigma^2/2)$, and hence find $\mathrm{E}(Y)$ and $\mathrm{var}(Y)$.
If $Y_1, \ldots, Y_n$ is a log-normal random sample, show that both $T_1 = \overline{Y}$ and $T_2 =$

$\exp(\overline{X} + S^2/2)$ are consistent estimators of $\mathrm{E}(Y)$, where $X_j = \log Y_j$ and S^2 is the sample variance of the X_j. Give the corresponding estimators of $\mathrm{var}(Y)$. Are the estimators based on the Y_j or on the X_j preferable? Why?

10 The binomial distribution models the number of 'successes' among independent variables with two outcomes such as success/failure or white/black. The multinomial distribution extends this to p possible outcomes, for example total failure/failure/success or white/black/red/blue/.... That is, each of the discrete variables $X_1, \ldots, X_m$ takes values $1, \ldots, p$, independently with probability $\Pr(X_j = r) = \pi_r$, where $\sum \pi_r = 1$, $\pi_r \geq 0$. Let $Y_r = \sum_j I(X_j = r)$ be the number of X_j that fall into category r, for $r = 1, \ldots, p$, and consider the distribution of $(Y_1, \ldots, Y_p)$.
(a) Show that the marginal distribution of Y_r is binomial with probability π_r, and that $\mathrm{cov}(Y_r, Y_s) = -m\pi_r\pi_s$, for $r \neq s$. Is it surprising that the covariance is negative?
(b) Hence give consistent estimators of positive probabilities π_r. What happens if some $\pi_r = 0$?
(d) Suppose that $p = 4$ with $\pi_1 = (2 + \theta)/4$, $\pi_2 = (1 - \theta)4$, $\pi_3 = (1 - \theta)/4$ and $\pi_4 = \theta/4$. Show that $T = m^{-1}(Y_1 + Y_4 - Y_2 - Y_3)$ is such that $\mathrm{E}(T) = \theta$ and $\mathrm{var}(T) = a/m$ for some $a > 0$. Hence deduce that T is consistent for θ as $m \to \infty$.
Give the value of T and its estimated variance when (y_1, y_2, y_3, y_4) equals $(125, 18, 20, 34)$.

2.3 Order Statistics

Summary statistics such as the sample median, interquartile range, and median absolute deviation are based on the ordered values of a sample $y_1, \ldots, y_n$, and they are also useful in assessing how closely a sample matches a specified distribution. In this section we study properties of ordered random samples.

The rth *order statistic* of a random sample $Y_1, \ldots, Y_n$ is $Y_{(r)}$, where

$$Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n-1)} \leq Y_{(n)}$$

is the ordered sample. We assume that the cumulative distribution F of the Y_j is continuous, so $Y_{(r)} < Y_{(r+1)}$ with probability one for each r and there are no ties.

Density function

The dy is a rhetorical device so that we can say the *probability* that $Y = y$ is $f(y)dy$.

To find the probability density of $Y_{(r)}$, we argue heuristically. Divide the line into three intervals: $(-\infty, y), [y, y + dy)$, and $[y + dy, \infty)$. The probabilities that a single observation falls into each of these intervals are $F(y), f(y)dy$, and $1 - F(y)$ respectively. Therefore the probability that $Y_{(r)} = y$ is

$$\frac{n!}{(r-1)!\ 1!\ (n-r)!} \times F(y)^{r-1} \times f(y)dy \times \{1 - F(y)\}^{n-r}, \tag{2.20}$$

where the second term is the probability that a prespecified $r-1$ of the Y_j fall in $(-\infty, y)$, the third the probability that a prespecified one falls in $[y, y+dy)$,

the fourth the probability that a prespecified $n-r$ fall in $[y+dy, \infty)$, and the first is a combinatorial multiplier giving the number of ways of prespecifying disjoint groups of sizes $r-1, 1$, and $n-r$ out of n.

If we drop the dy, expression (2.20) becomes a probability density function, from which we can derive properties of $Y_{(r)}$. For example, its mean is

$$\mathrm{E}(Y_{(r)}) = \frac{n!}{(r-1)!(n-r)!} \int_{-\infty}^{\infty} y f(y) F(y)^{r-1} \{1-F(y)\}^{n-r}\, dy \qquad (2.21)$$

when it exists; of course we expect that $\mathrm{E}(Y_{(1)}) < \cdots < \mathrm{E}(Y_{(n)})$.

Example 2.27 (Uniform distribution) Let $U_1, \ldots, U_n$ be a random sample from the uniform distribution on the unit interval,

$$\Pr(U \le u) = \begin{cases} 0, & u \le 0, \\ u, & 0 < u \le 1, \\ 1, & 1 < u; \end{cases} \qquad (2.22)$$

we write $U_1, \ldots, U_n \overset{\text{iid}}{\sim} U(0,1)$. As $f(u) = 1$ when $0 < u < 1$, $U_{(r)}$ has density

$$f_{U_{(r)}}(u) = \frac{n!}{(r-1)!(n-r)!} u^{r-1}(1-u)^{n-r}, \quad 0 < u < 1, \qquad (2.23)$$

and (2.21) shows that $\mathrm{E}(U_{(r)})$ equals

$$\begin{aligned} \frac{n!}{(r-1)!(n-r)!} \int_0^1 u\, u^{r-1}(1-u)^{n-r}\, dy &= \frac{n!}{(r-1)!(n-r)!} \frac{r!(n-r)!}{(n+1)!} \\ &= \frac{r}{n+1}; \end{aligned}$$

the value of the integral follows because (2.23) must have integral one for any r in the range $1, \ldots, n$ and any positive integer n. The expected positions of the n order statistics divide the unit interval and hence the total probability under the density into $n+1$ equal parts.

It is an exercise to show that $U_{(r)}$ has variance $r(n-r+1)/\{(n+1)^2(n+2)\}$ (Exercise 2.3.1). For large n this is approximately $n^{-1}p(1-p)$, where $p = r/n$, and hence we can write $U_{(r)} = r/(n+1) + \{p(1-p)/n\}^{1/2}\varepsilon$, where ε is a random variable with mean zero and variance approximately one. ■

Integrals such as (2.21) are nasty, but a good approximation is often available. Let $U, U_1, \ldots, U_n \overset{\text{iid}}{\sim} U(0,1)$ and $F^{-1}(u) = \min\{y : F(y) \ge u\}$. Then

Recall that every distribution function is right-continuous.

$$\Pr\{F^{-1}(U) \le y\} = \Pr\{U \le F(y)\} = F(y),$$

which is the distribution function of Y. Hence $Y \overset{D}{=} F^{-1}(U)$; note that for continuous F the variable $F(Y)$ has the $U(0,1)$ distribution; $F(Y)$ is called the *probability integral transform* of Y. It follows that $F^{-1}(U_1), \ldots, F^{-1}(U_n)$ is a random sample from F and that the joint distributions of the order statistics

$Y_{(1)}, \ldots, Y_{(n)}$ and of $F^{-1}(U_{(1)}), \ldots, F^{-1}(U_{(n)})$ are the same; in fact this is true for general F. Consequently $\mathrm{E}(Y_{(r)}) = \mathrm{E}\{F^{-1}(U_{(r)})\}$. But Example 2.27 implies that $U_{(r)} \stackrel{D}{=} r/(n+1) + \{p(1-p)/n\}^{1/2}\varepsilon$, where ε is a random variable with mean zero and unit variance. If we apply the delta method with $h = F^{-1}$, we obtain

$$\mathrm{E}(Y_{(r)}) = \mathrm{E}\{F^{-1}(U_{(r)})\} \doteq F^{-1}\{\mathrm{E}(U_{(r)})\} = F^{-1}\{r/(n+1)\}. \qquad (2.24)$$

Hence the plotting positions $F^{-1}\{r/(n+1)\}$ are approximate expected order statistics, justifying their use in probability plots; see Section 2.1.4.

Several order statistics

The argument leading to (2.20) can be extended to the joint distribution of any collection of order statistics. For example, the probability that the maximum, $Y_{(n)}$, takes value v and that the minimum, $Y_{(1)}$, takes value u, is

$$\frac{n!}{1!(n-2)!1!} \times f(u)du \times \{F(v) - F(u)\}^{n-2} \times f(v)dv, \quad u < v,$$

and is zero otherwise. Similarly the joint density of all n order statistics is

$$f_{Y_{(1)},\ldots,Y_{(n)}}(y_1, \ldots, y_n) = n! f(y_1) \times \cdots \times f(y_n), \quad y_1 < \cdots < y_n. \qquad (2.25)$$

In principle one can use (2.25) to calculate other properties of the joint distribution of the $Y_{(r)}$, but this can be very tedious. Here is an elegant exception:

Example 2.28 (Exponential order statistics) Consider the order statistics of a random sample $Y_1, \ldots, Y_n$ from the exponential density with parameter $\lambda > 0$, for which $\Pr(Y > y) = e^{-\lambda y}$. Let $E_1, \ldots, E_n$ denote a random sample of standard exponential variables, with $\lambda = 1$. Thus $Y_j \stackrel{D}{=} E_j/\lambda$.

The reasoning uses two facts. First, the distribution function of $\min(Y_1, \ldots, Y_r)$ is

$$\begin{aligned} 1 - \Pr\{\min(Y_1, \ldots, Y_r) > y\} &= 1 - \Pr\{Y_1 > y, \ldots, Y_r > y\} \\ &= 1 - \Pr(Y_1 > y) \times \cdots \times \Pr(Y_r > y) \\ &= 1 - \exp(-r\lambda y); \end{aligned}$$

this is exponential with parameter $r\lambda$. Second, the exponential density has the *lack-of-memory property*

$$\Pr(Y - x > y \mid Y > x) = \frac{\Pr(Y > x + y)}{\Pr(Y > x)} = \frac{\exp\{-\lambda(x+y)\}}{\exp(-\lambda x)} = \exp(-\lambda y),$$

implying that given that $Y - x$ is positive, its distribution is the same as the original distribution of Y, whatever the value of x.

We now argue as follows. Since $Y_{(1)} = \min(Y_1, \ldots, Y_n)$, its distribution is exponential with parameter $n\lambda$: $Y_{(1)} \stackrel{D}{=} E_1/(n\lambda)$. Given $Y_{(1)}$, $n-1$ of the Y_j

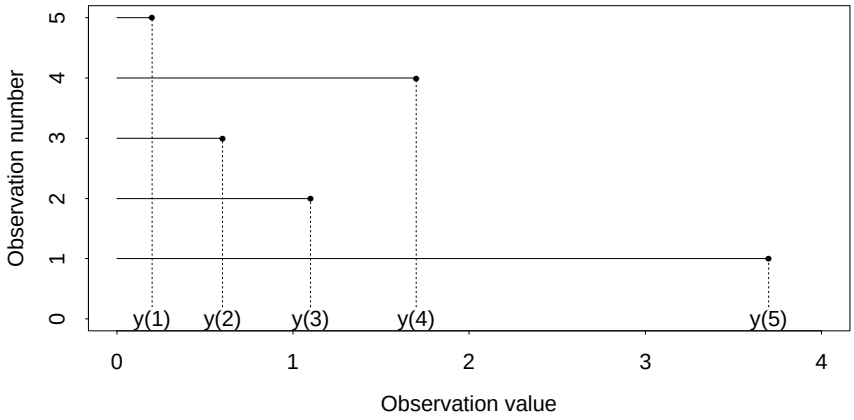

Figure 2.5 Exponential order statistics for a sample of size $n = 5$. The time to $y_{(1)}$ is the time to first event in a Poisson process of rate 5λ, and so it has the exponential distribution with mean $1/(5\lambda)$. The spacing $y_{(2)} - y_{(1)}$ is the time to first event in a Poisson process of rate 4λ, and is independent of $y_{(1)}$ because of the lack-of-memory property. It follows likewise that the spacings are independent and that the rth spacing has the exponential distribution with parameter $(n + 1 - r)\lambda$. During the second world war Alfréd Rényi (1921–1970) escaped from a labour camp and rescued his parents from the Budapest ghetto. He made major contributions to number theory and to probability. He was a gifted raconteur who defined a mathematician as 'a machine for turning coffee into theorems'.

remain, and by the lack-of-memory property the distribution of $Y_j - Y_{(1)}$ for each of them is the same as if the experiment had started at $Y_{(1)}$ with just $n-1$ variables; see Figure 2.5. Thus $Y_{(2)} - Y_{(1)}$ is exponential with parameter $(n-1)\lambda$, independent of $Y_{(1)}$, giving $Y_{(2)} - Y_{(1)} \stackrel{D}{=} E_2/\{(n-1)\lambda\}$. But given $Y_{(2)}$, just $n-2$ of the Y_j remain, and by the lack-of-memory property the distribution of $Y_j - Y_{(2)}$ for each of them is exponential independent of the past; hence $Y_{(3)} - Y_{(2)} \stackrel{D}{=} E_3/\{(n-2)\lambda\}$. This argument yields the *Rényi representation*

$$Y_{(r)} \stackrel{D}{=} \lambda^{-1} \sum_{j=1}^{r} \frac{E_j}{n+1-j}, \tag{2.26}$$

from which properties of the $Y_{(r)}$ are easily derived. For example,

$$\mathrm{E}\left(Y_{(r)}\right) = \lambda^{-1} \sum_{j=1}^{r} \frac{1}{n+1-j}, \ \mathrm{cov}\left(Y_{(r)}, Y_{(s)}\right) = \lambda^{-2} \sum_{j=1}^{r} \frac{1}{(n+1-j)^2}, \ s \geq r.$$

The upper right panel of Figure 2.3 shows a plot of the ordered times in the delivery suite against standard exponential plotting positions or *exponential scores*, $\sum_{j=1}^{r}(n+1-j)^{-1} \doteq -\log\{1 - r/(n+1)\}$. The exponential model fits very poorly.

The argument leading to (2.26) may be phrased in terms of Poisson processes. A superposition of independent Poisson processes is itself a Poisson process with rate the sum of the individual rates, so the period from zero to $Y_{(1)}$ is the time to the first event in a Poisson process of rate $n\lambda$, the time from $Y_{(1)}$ to $Y_{(2)}$ is the time to first event in a Poisson process of rate $(n-1)\lambda$, and

so on, with the times between events independent by definition of a Poisson process; see Figure 2.5. Exercise 2.3.4 gives another derivation. ■

Approximate density

Although (2.20) gives the exact density of an order statistic for a random sample of any size, approximate results are usually more convenient in practice. Suppose that r is the smallest integer greater than or equal to np, $r = \lceil np \rceil$, for some p in the range $0 < p < 1$. Then provided that $f\{F^{-1}(p)\} > 0$, we prove at the end of this section that $Y_{(r)}$ has an approximate normal distribution with mean $F^{-1}(p)$ and variance $n^{-1}p(1-p)/f\{F^{-1}(p)\}^2$ as $n \to \infty$. More formally,

$$\sqrt{n}\frac{\{Y_{(r)} - F^{-1}(p)\}f\{F^{-1}(p)\}}{\{p(1-p)\}^{1/2}} \xrightarrow{D} Z \quad \text{as} \quad n \to \infty, \tag{2.27}$$

where Z has a standard normal distribution.

Example 2.29 (Normal median) Suppose that $Y_1, \ldots, Y_n$ is a random sample from the $N(\mu, \sigma^2)$ distribution, and that $n = 2m+1$ is odd. The *median* of the sample is its central order statistic, $Y_{(m+1)}$. To find its approximate distribution in large samples, note that $(m+1)/(2m+1) \doteq \frac{1}{2}$ for large m, and since the normal density is symmetric about μ, $F^{-1}(\frac{1}{2}) = \mu$. Moreover $f(y) = (2\pi\sigma^2)^{-1/2}\exp\{-(y-\mu)^2/2\sigma^2\}$, so $f\{F^{-1}(\frac{1}{2})\} = (2\pi\sigma^2)^{-1/2}$. Thus (2.27) implies that in large samples $Y_{(m+1)}$ is approximately normal with mean μ and variance $\pi\sigma^2/(2n)$. ■

Example 2.30 (Birth data) In Figure 2.1 and Example 2.8 we saw that the daily medians of the birth data were generally smaller but more variable than the daily averages. To understand why, suppose that we have a sample of $n = 13$ observations from the gamma distribution F with mean $\mu = 8$ and shape parameter $\kappa = 3$; these are close to the values for the data. Then the average $\overline{Y}$ has mean μ and variance $\mu^2/(n\kappa)$; these are 8 and 1.64, comparable with the data values 7.90 and 1.54. The sample median has approximate expected value $F^{-1}(\frac{1}{2}) = 7.13$ and variance $n^{-1}\frac{1}{2}(1-\frac{1}{2})/f\{F^{-1}(\frac{1}{2})\}^2 = 4.02$, where f denotes the density (2.8); these values are to be compared with the average and variance of the daily medians, 7.03 and 2.15. The expected values are close, but the variances are not; we should not rely on an asymptotic approximation when $n = 13$. The theoretical variance of the median exceeds that of the average, so the sampling properties of the daily average and median are roughly what we might have expected: $\text{var}(M) > \text{var}(\overline{Y})$, and $\text{E}(M) < \text{E}(\overline{Y})$. Our calculation presupposes constant n, but in the data n changes daily; this is one source of error in the asymptotic approximation. ■

Expression (2.27) gives asymptotic distributions for *central order statistics*, i.e. $Y_{(r)}$ where $r/n \to p$ and $0 < p < 1$; as $n \to \infty$ such order statistics

have increasingly more values on each side. Different limits arise for *extreme order statistics* such as the minimum, for which $r = 1$ and $r/n \to 0$, and the maximum, for which $r = n$ and $r/n \to 1$. We discuss these more fully in Section 6.5.2, but here is a simple example.

Example 2.31 (Pareto distribution) Suppose that $Y_1, \ldots, Y_n$ is a random sample from the Pareto distribution, whose distribution function is

$$F(y) = \begin{cases} 0, & y < a, \\ 1 - (y/a)^{-\gamma}, & y \geq a, \end{cases}$$

where $a, \gamma > 0$. The minimum $Y_{(1)}$ exceeds y if and only if all the $Y_1, \ldots, Y_n$ exceed y, so $\Pr(Y_{(1)} > y) = (y/a)^{-n\gamma}$. To obtain a non-degenerate limiting distribution, consider $M = \gamma n(Y_{(1)} - a)/a$. Now

$$\Pr(M > z) = \Pr\left(Y_{(1)} > \frac{az}{n\gamma} + a\right) = \left(\frac{\frac{az}{n\gamma} + a}{a}\right)^{-n\gamma} \to e^{-z}$$

as $n \to \infty$. Consequently $\gamma n(Y_{(1)} - a)/a$ converges in distribution to the standard exponential distribution.

Vilfredo Pareto (1848–1923) studied mathematics and physics at Turin, and then became an engineer and director of a railway, before becoming professor of political economy in Lausanne. He pioneered sociology and the use of mathematics in economic problems. The Pareto distributions were developed by him to explain the spread of wealth in society.

There are two differences between this result and (2.27). First, and most obviously, the limiting distribution is not normal. Second, as the power of n by which $Y_{(1)} - a$ must be multiplied to obtain a non-degenerate limit is higher than in (2.27), the rate of convergence to the limit is faster than for central order statistics. Accelerated convergence of extreme order statistics does not always occur, however; see Example 6.32. ■

Derivation of (2.27)

This may be omitted at a first reading.

Consider $Y_{(r)}$, where $r = \lceil np \rceil$ and $0 < p < 1$ is fixed; hence $r/n \to p$ as $n \to \infty$. We saw earlier that $Y_{(r)} \stackrel{D}{=} F^{-1}(U_{(r)})$, where $U_{(r)}$ is the rth order statistic of a random sample $U_1, \ldots, U_n$ from the $U(0,1)$ density, and that $U_{(r)} = r/(n+1) + \{p(1-p)/n\}^{1/2}\varepsilon$, where ε has mean zero and variance tending to one as $n \to \infty$. Recall that F is a distribution whose density f exists. Hence the delta method gives $\mathrm{E}(Y_{(r)}) \doteq F^{-1}\{r/(n+1)\} \doteq F^{-1}(p)$, and as

$$\mathrm{var}\left(Y_{(r)}\right) = \mathrm{var}\{F^{-1}(U_{(r)})\} \doteq \mathrm{var}\left(U_{(r)}\right) \times \left\{\frac{dF^{-1}(p)}{dp}\right\}^2$$

and

$$\frac{d}{dp}F\{F^{-1}(p)\} = f\left\{F^{-1}(p)\right\}\frac{d}{dp}F^{-1}(p) = 1,$$

we have $\mathrm{var}\{Y_{(r)}\} \doteq p(1-p)/[f\{F^{-1}(p)\}^2 n]$ provided $f\{F^{-1}(p)\} > 0$.

To find the limiting distribution of $Y_{(r)}$, note that

$$\Pr(Y_{(r)} \le y) = \Pr\left(\sum_j I_j(y) \ge r\right), \tag{2.28}$$

where $I_j(y)$ is the indicator of the event $Y_j \le y$. The $I_j(y)$ are independent, so their sum $\sum_j I_j(y)$ is binomial with probability $F(y)$ and denominator n. Therefore (2.28) and the central limit theorem imply that for large n,

$$\Pr(Y_{(r)} \le y) \doteq 1 - \Phi\left(\frac{r - nF(y)}{[nF(y)\{1-F(y)\}]^{1/2}}\right). \tag{2.29}$$

Now choose $y = F^{-1}(p) + n^{-1/2}z\{p(1-p)/f\{F^{-1}(p)\}^2\}^{1/2}$, so that

$$F(y) = p + n^{-1/2}z\{p(1-p)\}^{1/2} + o(n^{-1/2}),$$

and recall that $r = \lceil np \rceil \doteq np$. Then (2.28) and (2.29) imply that, as required,

$$\Pr\left(n^{1/2}\frac{\left\{Y_{(r)} - F^{-1}(p)\right\}}{\left\{p(1-p)/f\left\{F^{-1}(p)\right\}^2\right\}^{1/2}} \le z\right)$$

approximately equals

$$1 - \Phi\left[\frac{np - np - n^{1/2}z\left\{p(1-p)\right\}^{1/2}}{\left\{np(1-p)\right\}^{1/2}}\right] = 1 - \Phi(-z) = \Phi(z).$$

Exercises 2.3

1 If $U_{(1)} < \cdots < U_{(n)}$ are the order statistics of a $U(0,1)$ random sample, show that $\text{var}(U_{(r)}) = r(n-r+1)/\{(n+1)^2(n+2)\}$. Find $\text{cov}(U_{(r)}, U_{(s)})$, $r < s$ and hence show that $\text{corr}(U_{(r)}, U_{(s)}) \to 1$ for large n as $r \to s$.

2 Let $U_1, \ldots, U_{2m+1}$ be a random sample from the $U(0,1)$ distribution. Find the exact density of the median, $U_{(m+1)}$, and show that $U_{(m+1)} \dot{\sim} N\left\{\frac{1}{2}, (8m)^{-1}\right\}$ for large m.

3 Let the $X_1, \ldots, X_n$ be independent exponential variables with rates λ_j. Show that $Y = \min(X_1, \ldots, X_n)$ is also exponential, with rate $\lambda_1 + \cdots + \lambda_n$, and that $\Pr(Y = X_j) = \lambda_j/(\lambda_1 + \cdots + \lambda_n)$.

4 Verify that the joint distribution of all the order statistics of a sample of size n from a continuous distribution with density $f(y)$ is (2.25). Hence find the joint density of the *spacings*, $S_1 = Y_{(1)}, S_2 = Y_{(2)} - Y_{(1)}, \ldots, S_n = Y_{(n)} - Y_{(n-1)}$, when $f(y) = \lambda e^{-\lambda y}$, $y > 0$, $\lambda > 0$. Use this to establish (2.26).

5 Use (2.27) to show that $Y_{(r)} \xrightarrow{P} F^{-1}(p)$ as $n \to \infty$, where $r = \lceil pn \rceil$ and $0 < p < 1$ is constant.

Consider IQR and MAD (Example 2.2). Show that IQR $\xrightarrow{P} 1.35\sigma$ for normal data and hence give an estimator of σ. Find also the estimator based on MAD.

6 Let N be a random variable taking values $0, 1, \ldots$, let $G(u)$ be the probability-generating function of N, let $X_1, X_2, \ldots$ be independent variables each having distribution function F, and let $Y = \max\{X_1, \ldots, X_N\}$. Show that Y has distribution function $G\{F(y)\}$, and find this when N is Poisson and the X_j exponential.

7 Let M and IQR be the median and interquartile range of a random sample $Y_1, \ldots, Y_n$ from a density of form $\tau^{-1}g\{(y-\eta)/\tau\}$, where $g(u)$ is symmetric about $u = 0$ and $g(0) > 0$. Show that as $n \to \infty$,

$$n^{1/2}\frac{M-\eta}{\text{IQR}} \xrightarrow{D} N(0, c),$$

for some $c > 0$, and give c in terms of g and its integral G.
Give c when $g(u)$ equals $\frac{1}{2}\exp(-|u|)$ and $\exp(u)/\{1+\exp(u)\}^2$.

8 The probability that events in a Poisson process of rate $\lambda > 0$ observed over the interval $(0, t_0)$ occur at $0 < t_1 < t_2 < \cdots < t_n < t_0$ is

$$\lambda^n \exp(-\lambda t_0), \quad 0 < t_1 < t_2 < \cdots < t_n < t_0.$$

By integration over $t_1, \ldots, t_n$, show that the probability that n events occur, regardless of their positions, is

$$\frac{(\lambda t_0)^n}{n!}\exp(-\lambda t_0), \quad n = 0, 1, \ldots,$$

and deduce that given that n events occur, the conditional density of their times is $n!/t_0^n$, $0 < t_1 < t_2 < \cdots < t_n < t_0$. Hence show that the times may be considered to be order statistics from a random sample of size n from the uniform distribution on $(0, t_0)$.

9 Find the exact density of the median M of a random sample $Y_1, \ldots, Y_{2m+1}$ from the uniform density on the interval $(\theta - \frac{1}{2}, \theta + \frac{1}{2})$. Deduce that $Z = m^{1/2}(M-\theta)$ has density

$$f(z) = \frac{(2m+1)!}{(m!)^2 m^{1/2}}\left(\frac{1}{4} + \frac{z^2}{m}\right)^m, \quad |z| < \tfrac{1}{2}m^{1/2},$$

and by considering the behaviour of $\log f(z)$ as $m \to \infty$ or otherwise, show that for large m, $Z \stackrel{\cdot}{\sim} N(0, 1/8)$. Check that this agrees with the general formula for the asymptotic distribution of a central order statistic.

Stirling's formula implies that $\log m! \sim \frac{1}{2}\log(2\pi) + (m+\frac{1}{2})\log m - m$ as $m \to \infty$.

2.4 Moments and Cumulants

Calculations involving moments often arise in statistics, but they are generally simpler when expressed in terms of equivalent quantities known as cumulants.

The *moment-generating function* of the random variable Y is $M(t) = \mathrm{E}(e^{tY})$, provided $M(t) < \infty$. Let

$$M'(t) = \frac{dM(t)}{dt}, \quad M''(t) = \frac{d^2M(t)}{dt^2}, \quad M^{(r)}(t) = \frac{d^rM(t)}{dt^r}, \quad r = 3, \ldots,$$

denote derivatives of M. If finite, the rth *moment* of Y is $\mu'_r = M^{(r)}(0) = \mathrm{E}(Y^r)$, giving the power series expansion

$$M(t) = \sum_{r=0}^{\infty} \mu'_r t^r / r!.$$

The characteristic function $\mathrm{E}(e^{itY})$, with $i^2 = -1$ is defined more broadly than $M(t)$, but as we shall not need the extra generality, $M(t)$ is used almost everywhere in this book.

The quantity μ'_r is sometimes called the rth moment about the origin, whereas $\mu_r = \mathrm{E}\{(Y - \mu'_1)^r\}$ is the rth *moment about the mean*. Among elementary properties of the moment-generating function are the following: $M(0) = 1$; the mean and variance of Y may be written

$$\mathrm{E}(Y) = M'(0),\ \mathrm{var}(Y) = M''(O) - \{M'(0)\}^2;$$

random variables $Y_1, \ldots, Y_n$ are independent if and only if their joint moment-generating function factorizes as

$$\mathrm{E}\{\exp(Y_1 t_1 + \cdots + Y_n t_n)\} = \mathrm{E}\{\exp(Y_1 t_1)\} \cdots \mathrm{E}\{\exp(Y_n t_n)\};$$

and the fact that any moment-generating function corresponds to a unique probability distribution.

Cumulants

The *cumulant-generating function* or *cumulant generator* of Y is the function $K(t) = \log M(t)$, and the rth *cumulant* is $\kappa_r = K^{(r)}(0) = d^r K(0)/dt^r$, giving the power series expansion

$$K(t) = \sum_{r=1}^{\infty} t^r \kappa_r / r!, \tag{2.30}$$

provided all the cumulants exist. Differentiation of (2.30) shows that the mean and variance of Y are its first two cumulants

$$\kappa_1 = K'(0) = \frac{M'(0)}{M(0)} = \mu'_1,\ \kappa_2 = K''(0) = \frac{M''(0)}{M(0)} - \frac{M'(0)^2}{M(0)^2} = \mu'_2 - (\mu'_1)^2.$$

Further differentiation gives higher-order cumulants. Cumulants are mathematically equivalent to moments, and can be defined as combinations of powers of moments, but we shall see below that their statistical interpretation is much more natural than is that of moments.

Example 2.32 (Normal distribution) If Y has the $N(\mu, \sigma^2)$ distribution, its moment-generating function is $M(t) = \exp(t\mu + \frac{1}{2}t^2\sigma^2)$ and its cumulant-generating function is $K(t) = t\mu + \frac{1}{2}t^2\sigma^2$. The first two cumulants are μ and σ^2, and all its higher-order cumulants are zero. The standard normal distribution has $K(t) = \frac{1}{2}t^2$. ■

The cumulant-generating function is very convenient for statistical work. Consider independent random variables $Y_1, \ldots, Y_n$ with respective cumulant-generating functions $K_1(t), \ldots, K_n(t)$. Their sum $Y_1 + \cdots + Y_n$ has cumulant-generating function

$$\log M_{Y_1+\cdots+Y_n}(t) = \log \mathrm{E}\left\{\exp(tY_1 + \cdots + tY_n)\right\} = \log \prod_{j=1}^{n} M_{Y_j}(t) = \sum_{j=1}^{n} K_j(t).$$

It follows that the rth cumulant of a sum of independent random variables is the sum of their rth cumulants. Similarly, the cumulant-generating function of a linear combination of independent random variables is

$$K_{a+\sum_{j=1}^{n} b_j Y_j}(t) = \log \mathrm{E}\left\{\exp(ta + tb_1Y_1 + \cdots + tb_nY_n)\right\} = ta + \sum_{j=1}^{n} K_j(b_j t). \tag{2.31}$$

Example 2.33 (Chi-squared distribution) If $Z_1, \ldots, Z_\nu$ are independent standard normal variables, each Z_j^2 has the chi-squared distribution on one degree of freedom, and (3.10) gives its moment-generating function, $(1 - 2t)^{-1/2}$. Therefore each Z_j^2 has cumulant-generating function $-\frac{1}{2}\log(1-2t)$, and the χ^2_ν random variable $W = \sum_{j=1}^{\nu} Z_j^2$ has cumulant-generating function

$$K(t) = -\frac{\nu}{2}\log(1-2t) = -\frac{\nu}{2}\sum_{r=1}^{\infty}(-1)^{r-1}\frac{(-2t)^r}{r} = \nu\sum_{r=1}^{\infty}2^{r-1}(r-1)!\frac{t^r}{r!},$$

provided that $|t| < \frac{1}{2}$. Therefore W has rth cumulant $\kappa_r = \nu 2^{r-1}(r-1)!$. In particular, the mean and variance of W are ν and 2ν. ∎

Example 2.34 (Linear combination of normal variables) Let $L = a + \sum_{j=1}^{n} b_j Y_j$ be a linear combination of independent random variables, where Y_j has the normal distribution with mean μ_j and variance σ_j^2. Then L has cumulant-generating function

$$at + \sum_{j=1}^{n}\left\{(b_jt)\mu_j + \frac{1}{2}(b_jt)^2\sigma_j^2\right\} = t\left(a + \sum_{j=1}^{n} b_j\mu_j\right) + \frac{t^2}{2}\left(\sum_{j=1}^{n} b_j^2\sigma_j^2\right),$$

corresponding to a $N(a + \sum b_j\mu_j, \sum b_j^2\sigma_j^2)$ random variable. ∎

Skewness and kurtosis

The third and fourth cumulants of Y are called its *skewness*, κ_3, and *kurtosis*, κ_4. Example 2.32 showed that $\kappa_3 = \kappa_4 = 0$ for normal variables. This suggests that they be used to assess the closeness of a variable to normality. However, they are not invariant to changes in the scale of Y, and the *standardized skewness* $\kappa_3/\kappa_2^{3/2}$ and *standardized kurtosis* κ_4/κ_2^2 are used instead for this purpose; small values suggest that Y is close to normal.

Some authors define the kurtosis to be $\kappa_4 + 3\kappa_2^2$, in our notation.

The average $\overline{Y}$ of a random sample of observations, each with cumulant-generating function $K(t)$, has mean and variance κ_1 and $n^{-1}\kappa_2$. Expression (2.31) shows that the random variable $Z_n = n^{1/2}\kappa_2^{-1/2}(\overline{Y} - \kappa_1)$, which is asymptotically standard normal, has cumulant-generating function

$$nK\left(n^{-1/2}\kappa_2^{-1/2}t\right) - n^{1/2}\kappa_2^{-1/2}\kappa_1 t,$$

and this equals

$$n\left\{\frac{t}{n^{1/2}}\frac{\kappa_1}{\kappa_2^{1/2}} + \tfrac{1}{2}\frac{t^2}{n}\frac{\kappa_2}{\kappa_2} + \tfrac{1}{6}\frac{t^3}{n^{3/2}}\frac{\kappa_3}{\kappa_2^{3/2}} + \tfrac{1}{24}\frac{t^4}{n^2}\frac{\kappa_4}{\kappa_2^2} + o\left(\frac{t^4}{n^2}\right)\right\} - n^{1/2}t\frac{\kappa_1}{\kappa_2^{1/2}}.$$

After simplification we find that the cumulant-generating function of Z_n is

$$\tfrac{1}{2}t^2 + \tfrac{1}{3}n^{-1/2}t^3\frac{\kappa_3}{\kappa_2^{3/2}} + \tfrac{1}{24}n^{-1}t^4\frac{\kappa_4}{\kappa_2^2} + o\left(\frac{t^4}{n}\right). \tag{2.32}$$

Hence convergence of the cumulant-generating function of Z_n to $\frac{1}{2}t^2$ as $n \to \infty$ is controlled by the standardized skewness and kurtosis $\kappa_3/\kappa_2^{3/2}$ and κ_4/κ_2^2.

Example 2.35 (Poisson distribution) Let $Y_1, \ldots, Y_n$ be independent Poisson observations with means $\mu_1, \ldots, \mu_n$. The moment-generating function of Y_j is $\exp\{\mu_j(e^t - 1)\}$, so its cumulant-generating function is $K_j(t) = \mu_j(e^t - 1)$ and all its cumulants equal μ_j. As the cumulant-generating function of $Y_1 + \cdots + Y_n$ is $\sum_j \mu_j(e^t - 1)$, the sum $\sum Y_j$ has a Poisson distribution with mean $\sum \mu_j$.

Now suppose that all the μ_j equal μ, say. From (2.31), the cumulant-generating function of the standardized average, $n^{1/2}\mu^{-1/2}(\overline{Y} - \mu)$, is

$$\begin{aligned} nK\left\{t(n\mu)^{-1/2}\right\} - t(n\mu)^{1/2} &= n\mu\left\{e^{t(n\mu)^{-1/2}} - 1\right\} - t(n\mu)^{1/2} \\ &= n\mu\sum_{r=2}^{\infty}\frac{t^r}{(n\mu)^{r/2}r!}. \end{aligned}$$

Thus $\overline{Y}$ has standardized skewness and kurtosis $(n\mu)^{-1/2}$ and $(n\mu)^{-1}$; in general $\kappa_r = (n\mu)^{-(r-2)/2}$ for $r = 2, 3, \ldots$ Hence $\overline{Y}$ approaches normality for fixed μ and large n or fixed n and large μ. ■

Vector case

A vector random variable $Y = (Y_1, \ldots, Y_p)^{\mathrm{T}}$ has moment-generating function $M(t) = \mathrm{E}(e^{t^{\mathrm{T}}Y})$, where $t^{\mathrm{T}} = (t_1, \ldots, t_p)$. The joint moments of the Y_r are the derivatives

$$\mathrm{E}(Y_1^{r_1}\cdots Y_p^{r_p}) = \left.\frac{\partial^{r_1+\cdots+r_p}M(t)}{\partial t_1^{r_1}\cdots\partial t_p^{r_p}}\right|_{t=0}.$$

The cumulant-generating function is again $K(t) = \log M(t)$, and the joint cumulants of the Y_r are given by mixed partial derivatives of $K(t)$ with respect

to the elements of t. For example, the covariance matrix of Y is the $p \times p$ symmetric matrix whose (r, s) element is $\kappa_{r,s} = \partial^2 K(t)/\partial t_r \partial t_s$, evaluated at $t = 0$.

Suppose that $Y = (Y_1, Y_2)^{\mathrm{T}}$, and that the scalar random variables Y_1 and Y_2 are independent. Then their joint cumulant-generating function is

$$K(t) = \log \mathrm{E}\left\{\exp(t_1 Y_1 + t_2 Y_2)\right\} = \log \mathrm{E}\left\{\exp(t_1 Y_1)\right\} + \log \mathrm{E}\left\{\exp(t_2 Y_2)\right\},$$

because the moment-generating function of independent variables factorizes. But since every mixed derivative of $K(t)$ equals zero, all the joint cumulants of Y_1 and Y_2 equal zero also. This observation generalizes to several variables: the joint cumulants of independent random variables are all zero. This is not true for moments, and partly explains why cumulants are important in statistical work.

Joint derivatives are not needed to obtain first cumulants, which are not joint cumulants.

Example 2.36 (Multinomial distribution) The probability density of a multinomial random variable $Y = (Y_1, \ldots, Y_p)^{\mathrm{T}}$ with denominator m and probabilities $\pi = (\pi_1, \ldots, \pi_p)$, that is $\Pr(Y_1 = y_1, \ldots, Y_p = y_p)$, equals

$$\frac{m!}{y_1! \cdots y_p!} \pi_1^{y_1} \cdots \pi_p^{y_p}, \quad y_r = 0, 1, \ldots, m, \quad \sum_{r=1}^{p} y_r = m;$$

note that $\pi_r \geq 0$, $\sum_r \pi_r = 1$. This arises when m independent observations take values in one of p categories, each falling into the rth category with probability π_r. Then Y_r is the total number falling into the rth category. If $Y_1, \ldots, Y_p$ are independent Poisson variables with means $\mu_1, \ldots, \mu_p$, then their joint distribution conditional on $Y_1 + \cdots + Y_p = m$ is multinomial with denominator m and probabilities $\pi_r = \mu_r / \sum \mu_s$.

The moment-generating function of Y is

$$\mathrm{E}\left(e^{t^{\mathrm{T}} Y}\right) = \sum \frac{m!}{y_1! \cdots y_p!} \pi_1^{y_1} \cdots \pi_p^{y_p} e^{y_1 t_1 + \cdots + y_p t_p} = \left(\pi_1 e^{t_1} + \cdots + \pi_p e^{t_p}\right)^m;$$

the sum is over all vectors $(y_1, \ldots, y_p)^{\mathrm{T}}$ of non-negative integers such that $\sum_r y_r = m$. Thus $K(t) = m \log\left(\pi_1 e^{t_1} + \cdots + \pi_p e^{t_p}\right)$. It follows that the joint cumulants of the elements of Y are

$$\begin{aligned}
\kappa_r &= m\pi_r, \\
\kappa_{r,s} &= m\left(\pi_r \delta_{rs} - \pi_r \pi_s\right), \\
\kappa_{r,s,t} &= m\left(\pi_r \delta_{rst} - \pi_r \pi_s \delta_{rt}[3] + 2\pi_r \pi_s \pi_t\right), \\
\kappa_{r,s,t,u} &= m\left\{\pi_r \delta_{rstu} - \pi_r \pi_s \left(\delta_{rt}\delta_{su}[3] + \delta_{stu}[4]\right) + 2\pi_r \pi_s \pi_t \delta_{ru}[6] - 6\pi_r \pi_s \pi_t \pi_u\right\};
\end{aligned}$$

here a Kronecker delta symbol such as δ_{rst} equals 1 if $r = s = t$ and 0 otherwise, and a term such as $\pi_r \pi_s \delta_{rt}[3]$ indicates $\pi_r \pi_s \delta_{rt} + \pi_s \pi_t \delta_{rs} + \pi_r \pi_t \delta_{st}$. The value of $\kappa_{r,s}$ implies that components of Y are negatively correlated,

because a large value for one entails low values for the rest. Zero covariance occurs only if $\pi_r = 0$, in which case Y_r is constant. ■

Exercises 2.4

1 Show that the third and fourth cumulants of a scalar random variable in terms of its moments are

$$\kappa_3 = \mu_3' - 3\mu_1'\mu_2' + 2(\mu_1')^3, \ \kappa_4 = \mu_4' - 4\mu_3'\mu_1' - 3(\mu_2')^2 + 12\mu_2'(\mu_1')^2 - 6(\mu_1')^4.$$

2 Show that the cumulant-generating function for the gamma density (2.7) is $-\kappa\log(1 - t/\lambda)$. Hence show that $\kappa_r = \kappa(r-1)!/\lambda^r$, and confirm the mean, variance, skewness and kurtosis in Examples 2.12 and 2.26.
If $Y_1, \ldots, Y_n$ are independent gamma variables with parameters $\kappa_1, \ldots, \kappa_n$ and the same λ, show that their sum has a gamma density, and give its parameters.

This demands nodding acquaintance with characteristic functions.

3 The Cauchy density (2.16) has no moment-generating function, but its characteristic function is $\mathrm{E}(e^{itY}) = \exp(it\theta - |t|)$, where $i^2 = -1$. Show that the average $\overline{Y}$ of a random sample $Y_1, \ldots, Y_n$ of such variables has the same characteristic function as Y_1. What does this imply?

2.5 Bibliographic Notes

The idea that variation observed around us can be represented using probability models provides much of the motivation for the study of probability theory and underpins the development of statistics. Cox (1990) and Lehmann (1990) give complementary general discussions of statistical modelling and a glance at any statistical library will reveal hordes of books on specific topics, references to some of which are given in subsequent chapters. Real data, however, typically refuse to conform to neat probabilistic formulations, and for useful statistical work it is essential to understand how the data arise. Initial data analysis typically involves visualising the observations in various ways, examining them for oddities, and intensive discussion to establish what the key issues of interest are. This requires creative lateral thinking, problem solving, and communication skills. Chatfield (1988) gives very useful discussion of this and related topics.

John Wilder Tukey (1915–2000) was educated at home and then studied chemistry and mathematics at Brown University before becoming interested in statistics during the 1939–45 war, at the end of which he joined Princeton University. He made important contributions to areas including time series, analysis of variance, and simultaneous inference. He underscored the importance of data analysis, computing, robustness, and interaction with other disciplines at a time when mathematical statistics had become somewhat introverted, and invented many statistical terms and

J. W. Tukey and his co-workers have played an important role in stimulating development of approaches to exploratory data analysis both numerical and graphical; see Tukey (1977), Mosteller and Tukey (1977), and Hoaglin *et al.* (1983, 1985, 1991). Two excellent books on statistical graphics are Cleveland (1993, 1994), while Tufte (1983, 1990) gives more general discussions of visualizing data. For a brief account see Cox (1978).

Cox and Snell (1981) give an excellent general account of applied statistics.

Most introductory texts on probability and random processes discuss the main convergence results; see for example Grimmett and Stirzaker (2001).

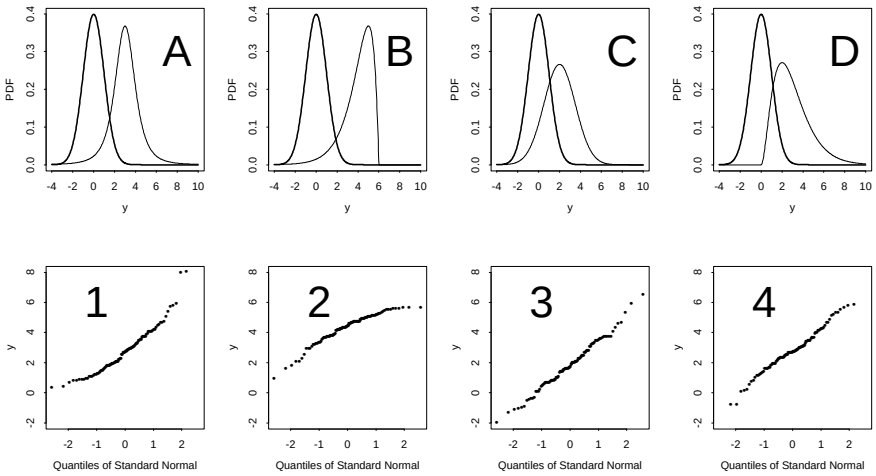

Figure 2.6 Match the sample to the density. Upper panels: four densities compared to the standard normal (heavy). Lower panels: normal probability plots for samples of size 100 from each density.

Bickel and Doksum (1977) give a more statistical account; see their page 461 for a proof of Slutsky's lemma. See also Knight (2000).

Arnold *et al.* (1992) give a full account of order statistics and many further references.

Most elementary statistics texts do not describe cumulants despite their usefulness. McCullagh (1987) contains forceful advocacy for them, including powerful methods for cumulant calculations. See also Kendall and Stuart (1977), whose companion volumes (Kendall and Stuart, 1973, 1976) overlap considerably with parts of this book, from a quite different viewpoint.

2.6 Problems

Pin the tail on the density.

1 Figure 2.6 shows normal probability plots for samples from four densities. Which goes with which?

2 Suppose that conditional on μ, X and Y are independent Poisson variables with means μ, but that μ is a realization of random variable with density $\lambda^\nu \mu^{\nu-1} e^{-\lambda\mu}/\Gamma(\nu)$, $\mu > 0$, $\nu, \lambda > 0$. Show that the joint moment-generating function of X and Y is

$$\mathrm{E}\left(e^{sX+tY}\right) = \lambda^\nu \left\{\lambda - (e^s - 1) - (e^t - 1)\right\}^{-\nu},$$

and hence find the mean and covariance matrix of (X, Y). What happens if $\lambda = \nu/\xi$ and $\nu \to \infty$?

3 Show that a binomial random variable R with denominator m and probability π has cumulant-generating function $K(t) = m \log\left(1 - \pi + \pi e^t\right)$. Find $\lim K(t)$ as $m \to \infty$ and $\pi \to 0$ in such a way that $m\pi \to \lambda > 0$. Show that

$$\Pr(R = r) \to \frac{\lambda^r}{r!} e^{-\lambda},$$

and hence establish that R converges in distribution to a Poisson random variable. This yields the Poisson approximation to the binomial distribution, sometimes called the *law of small numbers*. For a numerical check in the `S` language, try

```
y <- 0:10;  lambda <- 1; m <- 10; p <- lambda/m
round(cbind(y,pbinom(y,size=m,prob=p),ppois(y,lambda)),digits=3)
```

with various other values of m and λ.

4 (a) Let X be the number of trials up to and including the first success in a a sequence of independent Bernoulli trials having success probability π. Show that $\Pr(X = k) = \pi(1-\pi)^{k-1}$, $k = 1, 2, \ldots$, and deduce that X has moment-generating function $\pi e^t/\{1-(1-\pi)e^t\}$; hence find its mean and variance. X has the *geometric distribution*.
(b) Now let Y_n be the number of trials up to and including the nth success in such a sequence of trials. Show that

$$\Pr(Y_n = k) = \binom{k-1}{n-1}\pi^n(1-\pi)^{k-n}, \quad k = n, n+1, \ldots;$$

this is the *negative binomial distribution*. Find the mean and variance of Y_n, and show that as $n \to \infty$ the sequence $\{Y_n\}$ satisfies the conditions of the Central Limit Theorem. Deduce that

$$\lim_{n\to\infty} 2^{1-n}\sum_{k=0}^{n}\binom{k+n-1}{n-1}\frac{1}{2^k} = 1.$$

(c) Find the limiting cumulant-generating function of $\pi Y_n/(1-\pi)$ as $\pi \to 0$, and hence show that the limiting distribution is gamma.

5 Let $Y_1, \ldots, Y_n$ be a random sample from a distribution with mean μ and variance σ^2. Find the mean of

$$T = \frac{1}{2n(n-1)}\sum_{j\neq k}(Y_j - Y_k)^2,$$

and by writing $Y_j - Y_k = Y_j - \overline{Y} - (Y_k - \overline{Y})$, show that $T = S^2$.

6 Let $Y_1, \ldots, Y_n$ be a random sample from the uniform distribution on the interval $(\theta - \frac{1}{2}, \theta + \frac{1}{2})$. Show that the joint density of the sample maximum and minimum, $Y_{(n)}$ and $Y_{(1)}$, is

$$f_{Y_{(1)},Y_{(n)}}(u,v) = n(n-1)(v-u)^{n-2}, \quad \theta - \tfrac{1}{2} < u < v < \theta + \tfrac{1}{2}.$$

The *sample range* is $R = Y_{(n)} - Y_{(1)}$, and a natural estimator of θ is the *midrange*, $T = (Y_{(n)} + Y_{(1)})/2$. Show that the conditional density of T given R is

$$f(t \mid r; \theta) = (1-r)^{-1}, \quad 0 < r < 1, \ \theta + \tfrac{1}{2} - \frac{r}{2} > t > \theta - \tfrac{1}{2} + \frac{r}{2}.$$

How precisely is θ determined by this density as $r \to 0$ and $r \to 1$?

Waloddi Weibull (1887–1979) was a Swedish engineer who in 1937 published the distribution that bears his name; it is widely used in reliability.

7 A random variable X with the *Weibull distribution* with index α has distribution function $1 - \exp\{-(x/\lambda)^\alpha\}$, $x > 0$, $\lambda, \alpha > 0$. The idea that a system with many similar components will fail when the weakest component fails has led to widespread use of this distribution in industrial reliability.

(a) Suppose that $X_1, \ldots, X_n$ are independent identically distributed continuous non-negative random variables such that as $t \to 0$, the density and distribution functions are asymptotically $at^{\kappa-1}$ and at^α/α respectively, where $a, \alpha > 0$. Let $Y = \min(X_1, \ldots, X_n)$ and let $W = (a/\alpha)^{1/\alpha} n^{1/\alpha} Y$. Show that as $n \to \infty$, W has as its limiting distribution the Weibull distribution with index α.
(b) Explain why a probability plot for the Weibull distribution may be based on plotting the logarithm of the rth order statistic against $\log\left\{-\log\left(1 - \frac{r}{n+1}\right)\right\}$, and give the slope and intercept of such a plot. Check whether the data in Table 1.2 follow Weibull distributions.

8 Let $Y_1, \ldots, Y_{2m+1}$ be a random sample from the uniform density

$$f(y) = \begin{cases} \theta^{-1}, & 0 \le y \le \theta, \\ 0, & \text{otherwise.} \end{cases}$$

Derive the density function of the sample median $T = Y_{(m+1)}$ and find its exact mean and variance.
Find the density function of $Z = 2(2m+3)^{1/2}(Y_{(m+1)} - \theta/2)/\theta$ and use Stirling's formula to show directly that, as $m \to \infty$, Z has asymptotically a standard normal distribution. Deduce that asymptotically $\mathrm{var}(T) \sim 3\mathrm{var}(\overline{Y})$.

9 The *coefficient of variation* of a random sample $Y_1, \ldots, Y_n$ is $C = S/\overline{Y}$, where $\overline{Y}$ and S^2 are the sample average and variance. It estimates the ratio $\psi = \sigma/\mu$ of the standard deviation relative to the mean. Show that

$$\mathrm{E}(C) \doteq \psi, \quad \mathrm{var}(C) \doteq n^{-1}\left(\psi^4 - \gamma_3\psi^3 + \tfrac{1}{4}\gamma_4\psi^2\right) + \frac{\psi^2}{2(n-1)}.$$

10 If T_1 and T_2 are two competing estimators of a parameter θ, based on a random sample $Y_1, \ldots, Y_n$, the *asymptotic efficiency* of T_1 relative to T_2 is $\lim_{n\to\infty} \mathrm{var}(T_2)/\mathrm{var}(T_1) \times 100\%$. If $n = 2m + 1$, find the asymptotic efficiency of the sample median $Y_{(m+1)}$ relative to the average $\overline{Y} = n^{-1}\sum_j Y_j$ when the density of the Y_j is: (a) normal with mean θ and variance σ^2; (b) Laplace, $(2\sigma)^{-1}\exp\{-|y-\theta|/\sigma\}$ for $-\infty < y < \infty$; and (c) Cauchy, $\sigma/[\pi\{\sigma^2 + (y-\theta)^2\}]$ for $-\infty < y < \infty$.

11 Show that the covariance matrix for the multinomial distribution may be written $m(\mathrm{diag}\{\pi\} - \pi\pi^{\mathrm{T}})$, and deduce that it has determinant zero. Explain why the distribution is degenerate.

12 (a) If X has the $N(\mu, \sigma^2)$ distribution, show that X^2 has cumulant-generating function

$$t\mu^2/(1 - 2t\sigma^2) - \tfrac{1}{2}\log(1 - 2t\sigma^2).$$

(b) If $X_1, \ldots, X_\nu$ are independent normal variables with variance σ^2 and means $\mu_1, \ldots, \mu_\nu$, show that the cumulant-generating function of $W = X_1^2 + \cdots + X_\nu^2$ is

$$t\delta^2\sigma^2/(1 - 2t\sigma^2) - \tfrac{\nu}{2}\log(1 - 2t\sigma^2),$$

where $\delta^2 = (\mu_1^2 + \cdots + \mu_\nu^2)/\sigma^2$. The distribution of W/σ^2 is said to be *non-central chi-squared* with ν degrees of freedom and non-centrality parameter δ^2. Show that the moment-generating function of W may be written

$$\exp\left\{-\tfrac{1}{2}\delta^2 + \tfrac{1}{2}\delta^2(1 - 2t\sigma^2)^{-1}\right\}(1 - 2t\sigma^2)^{-\nu/2},$$

and that this equals

$$e^{-\delta^2/2} \sum_{r=0}^{\infty} \frac{1}{r!} \left(\frac{\delta^2}{2} \right)^r (1 - 2t\sigma^2)^{-r-\nu/2}. \tag{2.33}$$

Use (2.33) and (3.10) to write down an expression for the density of W.

(c) Hence deduce that (i) $W \stackrel{D}{=} W_\nu + W_{2N}$, where $W \sim \sigma^2 \chi^2_\nu$ independent of $W_{2N} \sim \sigma^2 \chi^2_{2N}$, with χ^2_0 taking value 0 with unit probability, and N is Poisson with mean $\delta^2/2$, and (ii) $W \stackrel{D}{=} (\delta\sigma + Y_1)^2 + Y_2^2 + \cdots + Y_\nu^2$.