Chapter 1
Probability, Statistics, and Reality

If your experiment needs statistics, you ought to have done a better experiment.—Ernest Rutherford (attrib.)

A branch of Science that must depend heavily on statistical inference is therefore in grave peril of error. Without the most rigorous and critical analysis of its concepts and methods it may fall to worshipping very strange gods.—John Ziman (1968) *Public Knowledge: an Essay Concerning the Social Dimension of Science* .

It is commonly only in their 4th or 5th [undergraduate] years that geophysics students first come into contact with statistical reality, usually in the form of some practical project or their MSc thesis. Then their fate is suddenly made manifest to them: they need statistics to analyze their data. Somehow, by hook, crook, or cookery book, they acquire a smattering of statistical jargon sufficient to appease the sadistic examiner with an eye for the missing confidence interval or misunderstood significance level. Small wonder if they thereafter view statisticians with reserve, as at best the curators of an elaborate maze through which they were compelled to wander, and at worst the cause of their downfall.—David Vere-Jones (1975), Stochastic models for earthquake sequences, *Geophys. J. R. Astron. Soc.***42***, 811-826.*

1. The Aim of the Exercise

The quotations above attempt to explain why this course exists. First of all, geophysics (like astronomy) is an observational science, in which we cannot do experiments—so Rutherford's dictum is unhelpful. We have to understand the magnetic field of the Earth as it is, without recourse to manipulation in the laboratory. While statistical methods are not confined to the observational science—nowadays even nuclear physicists use them[1]—they are so central to doing geophysics that they are a key part of a geophysicist's training.

The challenge of such training is that, as our second and third quotations indicate, statistics is difficult to learn and to do well. This difficulty is not just because of the mathematics involved, but also because we often reason poorly about random events, to the benefit of casinos and the frustration of students. We hope this course will help, though there is really no substitute for lots of experience, careful thought, and a willingness to subject one's intuition to careful checks.

In a one-quarter course we can only cover some aspects of the subject, though we hope to discuss much of what is needed in geophysics—which is not necessarily the same as the standard set of subjects in most introductory statistics books. We postpone to the second part of the course the very important case of data arranged in time or space. We neither seek nor shirk mathematics, though we only rarely attempt rigor; we shall show a number of theorems, but prove few. Familiarity with calculus, at least through multivariate calculus, and with vectors, vector spaces, and matrices, should be adequate mathematical background.

[1] Actually, this has been true for as long as there has been nuclear physics: for example, using the statistics of a Poisson process (which we will describe below) to show that radioactive decay occurs randomly. See E. Rutherford and H. Geiger, The probability variations in the distribution of alpha-particles, *Phil. Mag. Ser 6*, **20**, 698-707 (1910).

To give some of the flavor of statistical reasoning, we offer some examples of how it is applied—including one case of it being applied badly. While we use a few terms from probability and statistics in these examples, they are names that you are likely to already have some familiarity with; in any case we will define them more precisely in later chapters.

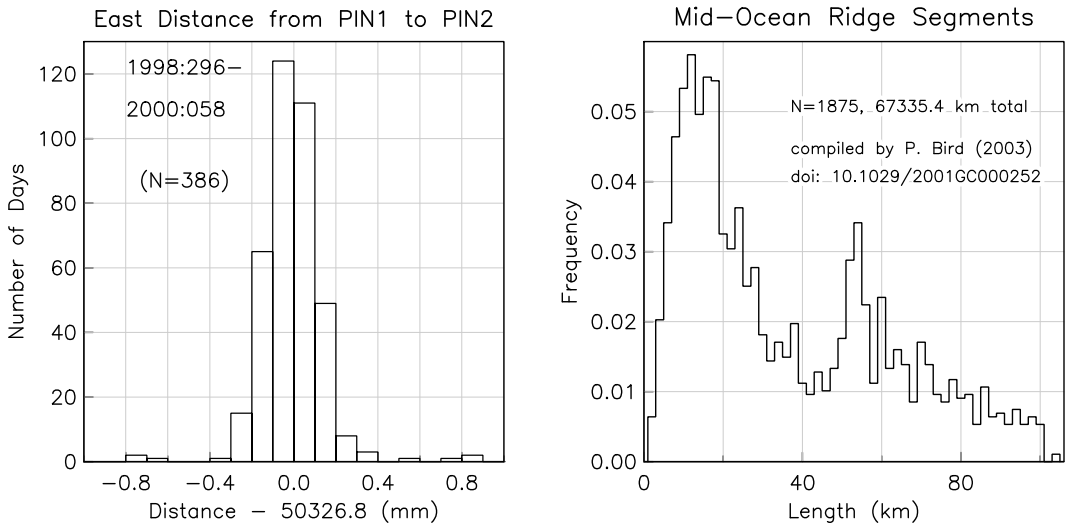

Figure 1.1

1.1. What Kind of Problem Are We Trying to Solve?

To offer something more concrete, we begin with a couple of examples of data, which show two different situations in which statistical reasoning can be needed. The left-hand plot in Figure 1.1 summarizes two years of measurements between two geodetic markers about 50 meters apart, using the satellite Global Positioning System (GPS). We plot these data in the familiar form of a histogram, which shows that the data tend to clump around a value of 50327 mm, but are spread around over a range of ±1 mm. In this case the usual frame for looking at the problem is what we might call the "physics-lab" one: there is a single, true value of the distance, and the reason that we get numbers that are scattered around this is because of errors in the measurements. Statistics then is just a necessary tool for dealing with something we would rather not exist in the first place.

This frame is fine in the laboratory, but often not useful outside it. For an example, consider the right-hand plot in Figure 1.1, showing a histogram of the lengths of segments of oceanic spreading centers (terminated by some other kind of plate boundary). In this case, there is no ideal "true" length, and also (at this scale) not much measurement error.[2] Rather, the way in which the numbers are distributed is the information, which we treat using statistical methods.

What is common to these examples is that we have numbers that are in some way scattered, or varying. We use probability theory, and the statistical methods that flow from it, because this a mathematical construct most suited to dealing with

[2] Though we may wonder, reasonably, about the effect of sampling: a long segment may be one that doesn't have many ship tracks crossing it.

such variations. More precisely, our aim, given a set of data, is to develop a mathematical model for it: such models are called probability models or **stochastic models**.

Table 1

2, 7	2, 10	5, 5	7, 4	7, 9	4, 8	10, 5	6, 8	7, 9	5, 7
8, 3	4, 9	8, 7	5, 6	5, 7	9, 7	6, 7	8, 11	3, 5	7, 6
6, 3	8, 5	10, 8	6, 5	3, 4	10, 3	9, 6	5, 6	10, 7	3, 5
9, 3	5, 8	4, 4	8, 7	6, 3	4, 6	5, 10	5, 8	6, 8	9, 8
8, 11	5, 5	8, 8	9, 7	8, 6	7, 8	10, 5	7, 5	8, 6	7, 6
2, 7	7, 7	9, 5	5, 4	12, 6	7, 6	2, 6	6, 5	6, 6	2, 9
9, 7	5, 5	8, 5	5, 6	8, 5	5, 6	4, 6	4, 3	8, 8	9, 7
7, 8	5, 5	6, 6	8, 8	7, 5	8, 7	10, 5	4, 6	9, 5	4, 5
6, 8	7, 5	2, 6	2, 5	4, 6	3, 8	6, 7	6, 5	7, 6	7, 4
5, 7	11, 4	10, 7	5, 5	4, 4	4, 8	7, 4	11, 5	7, 9	2, 6

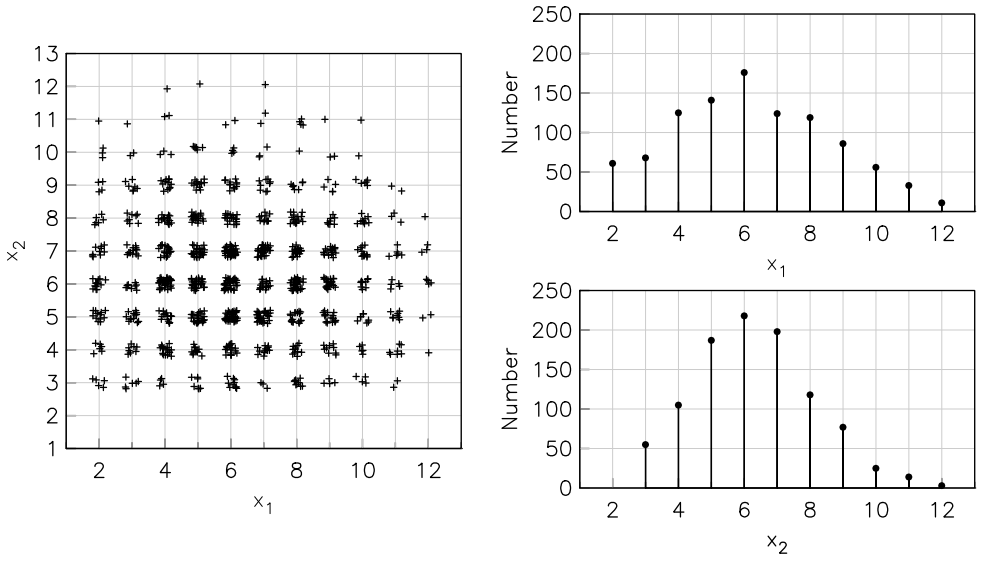

Figure 1.1

1.1.1. A Toy Example of a Stochastic Model

To show such a model, it is best to start with a made-up example. Table 1 shows 100 pairs of data x_1, x_2) that might be collected from repeating an experiment; clearly the numbers vary considerably, though (by design, in this case) they are all integers. The first thing to do with any dataset should be to plot it: humans are visual, not (naturally) numerical, and it is as well to make good use of this. Figure 1.1 shows a plot of 1000 data pairs. Since the values are all integers, we have moved each point away from the true value by a small random amount so we can see the relative number at each value, a plotting trick known as **dithering**. (We will discuss other good and bad plotting methods later on). The figure also shows the relative number of points with particular values of x_1 and x_2, in plots on the right. Clearly there is a pattern to the relative occurrences of the different values.

In fact, these data were generated in the following way: take three numbers, each with an equal chance of being 1, 2, 3, or 4, and add them up to get the value of

x_1; and take two numbers, each with an equal chance of being 1, 2, 3, 4, 5, or 6, and add them up to get the value of x_2. This could be done by rolling tetrahedral and cubical dice, though in this case it was done by a software simulation.

This brief description, which is enough to explain all these data, is an example of a stochastic model: a mathematical construction that we develop to explain some set of observations. This model includes, as a basic element, some aspects that are "random", meaning that they do not have definite values, but only different chances of taking on different values. In the model we have set out, we start with elements that have equal chances of taking on integer values within certain ranges, and combine them to get our final result. These elements are called **random variables**, and a good part of the application of stochastic models comes from deciding what parts of your data are best modeled by variables of this type, and which by variables which take on only one value. The mathematics of random variables is part of **probability theory**, which like any mathematical theory is the logical development of the behavior of certain kinds of mathematical entities. Random variables are such an entity, designed (as their name suggests) to mimic outcomes that in the real world are the result of "chance".

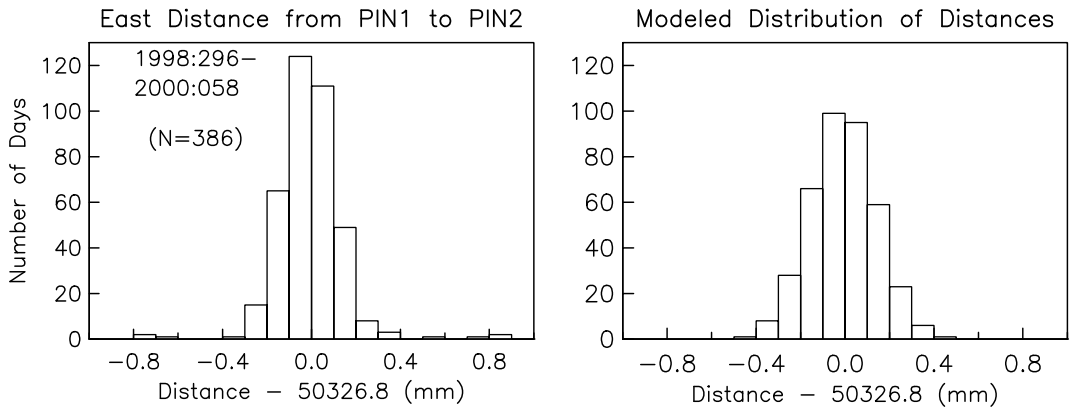

Figure 1.2

1.2. Distance Measurement (I): Point Estimation

We now return to the first example data set, shown again in histogram form on the left plot in Figure 1.2. To go beyond simply noting that the data are clumped around a value, we have to assume that we can represent the data through a mathematical model that includes random variables. such a stochastic model we are asserting that the data we have are at least in part the result of a random process. This is a large conceptual leap—and like most applications of mathematical models to the real world, not one that is simple to justify. You should always remember that assuming a relationship between data and a model that describes them should be done with care, for it is, in its own way, chancy: a stochastic model might or might not be appropriate.

The specific mathematical model we adopt is that the length can be modeled as a random variable X; this variable is described by the equation

$$\mathcal{P}(X < x) = \frac{1}{\sigma(2\pi)^{\frac{1}{2}}} \int_{-\infty}^{x} \exp\left[-\tfrac{1}{2}\left(\frac{u-m}{\sigma}\right)^2\right] du \tag{1}$$

The right-hand side of this equation should be familiar: it describes a function of x that depends on two parameters m and σ, which have definite (but at this point unknown) values. The left-hand side is where the novelty lies, both in notation and in concept. The expression means, "The probability that the random variable X is less than a value x."

But what do we mean by probability? We will describe its mathematical properties in the next chapter. What it "really means", in the sense of what it corresponds to in the real world, is a deep, and deeply controversial, question. A common "working definition" is that it is the fraction of times that we get a particular result if we repeat something: the frequency of occurrence, as in the canonical example that the probability of getting heads on tossing a fair coin is 0.5. While there are good objections to this interpretation, there are good objections to all the other ones as well. Certainly the frequency interpretation seems like the most straightforward way to connect equation (1) to the data summarized in Figure 1.2.

So, we have data, and an assumed probability model for them. What comes next is **statistics**: using the model to draw conclusions from the data. Probability theory is mathematics, with its own rules—albeit rules inspired by real-world examples. Statistics, while it makes use of probability theory, is the activity of applying it to actual data, and the drawing of actual conclusions. Statistics can perhaps best be thought of as a branch of applied mathematics, though for historical reasons nobody uses this term for it.

Given our model, as expressed by equation (1), and given these data, an immediate statistical question is, what are the "best" values for m and σ? Assuming that (1) in fact correctly models these data, the formulas for finding the "best" m and σ are

$$\hat{m} = \frac{1}{N}\sum_{i=1}^{N} d_i \tag{2}$$

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N} (d_i - \hat{m})^2$$

where there are N data $d_1, d_2, \cdots, d_N$. Applying these expressions to the data from which Figure 1.3 was drawn, we get $\hat{m} = 50326.795$ and $\hat{\sigma} = 0.18$. The parameters m and σ in equation (1) are called the **mean** and **standard deviation**; the values $\hat{m}$ and $\hat{\sigma}$ that we get from the data using the formulae in (2) are called **estimates** of these. (As is standard in statistics, we use a superposed hat symbol to denote an estimate of a variable). Why these estimates might be termed the "best" ones is something we will get to; this is part of the area of statistics known, unsurprisingly, as **estimation theory**. Finding the best values of parameters such as m and σ is called **point estimation**. As we describe in the next subsection, this is the answer to only one kind of statistical question; perhaps the commonest, but not always the most relevant.

The interpretation of these data as one of a true value contaminated by errors leads to a formally similar but philosophically quite different statement: that the data can be modeled as

$$d_i = t + e_i \tag{3}$$

where t is the true value of the distance, and e_i are the "errors"; we model the errors as being a random variable E, such that

$$\mathcal{P}(E < e) = \frac{1}{\sigma(2\pi)^{\frac{1}{2}}} \int_{-\infty}^{e} \exp\left[-\tfrac{1}{2}\left(\frac{u}{\sigma}\right)^2\right] du \tag{4}$$

Mathematically, (3) and (4) are equivalent to (1) if we have $t = m$, which means that if we assume (3) and (4) our "best estimate" of t is $\hat{m}$. This is of course the bit of statistics scientists do most often: given a collection of repeated measurements, we form the average, and say the answer is the true result we are trying to find. There is nothing really wrong in this—until we start to ask questions such as "How uncertain is t?" Such a question, while tempting, is nonsense: if t is the true value, it cannot be uncertain mathematically, it is some number, not (like the e's) a random kind of thing. We will therefore avoid this approach, and instead will try to pose problems in terms of equations such as (1), in which all the non-random variables are, as it were, on the same footing: they are all **parameters** in a formula describing a probability.

We can use our estimates $\hat{m}$ and $\hat{\sigma}$ to find how the data would be distributed if they actually followed the model (1); the right side of Figure 1.3 shows the result in the same histogram form as the left side. You might wish to argue, looking at these plots, that the model is in fact not right, because the data show a sharper peak, just to the left of zero, than the model; and the model does not show the few points at large distances that are evident in the data. As it turns out, these are quite valid objections, which lead, among other things, to something called **robust estimation**: how to make point estimates that are not affected by small amounts of outlying data. But the underlying question, of deciding if the model we are using is a valid one or not, is quite a different issue from estimating model parameters; so we discuss it in a new section.

1.3. Another Example: Hypothesis Testing

Figure 1.4 plots another dataset: the top two traces show the reversals of the Earth's magnetic field, as reconstructed mostly from marine magnetic anomalies, over the last 159 Myr. We plot this, as is conventional, as a kind of square wave, with positive values meaning that the dipole has the same direction as now, and negative ones that it was reversed. This is an example of a geophysical time series of a particular kind, namely a **point process**, so called because such a time series is defined by points in time at which something happens: a field reversal, an earthquake, or a disk crash. The simplest probability model for this is called a **Poisson process**: in any given unit of time (we suppose for simplicity that time is divided into units) there is a probability p that the event happens. By this specification, p is always the same at any date, and it does not matter if it has been a long time or a

short time since the last event. Such a model, which doesn't know what time it is, does not describe the actual times of the reversals. What it does describe is the intervals between successive events, which we plot on the lower left of Figure 1.2, again as a histogram. We use the log of the time because the intervals have to be positive—and also because they are spread out over a large range, from 2×10^4 to 4×10^7 years. The longest interval (the "Cretaceous Normal Superchron") ran from 124 Myr ago to 83 Myr ago. Looking at the time series, or the distribution, naturally raises a question: is this interval somehow "unusual" compared to the others? If it is, we might wish to argue that the core dynamo (the source of the field) changed its behavior during this time.

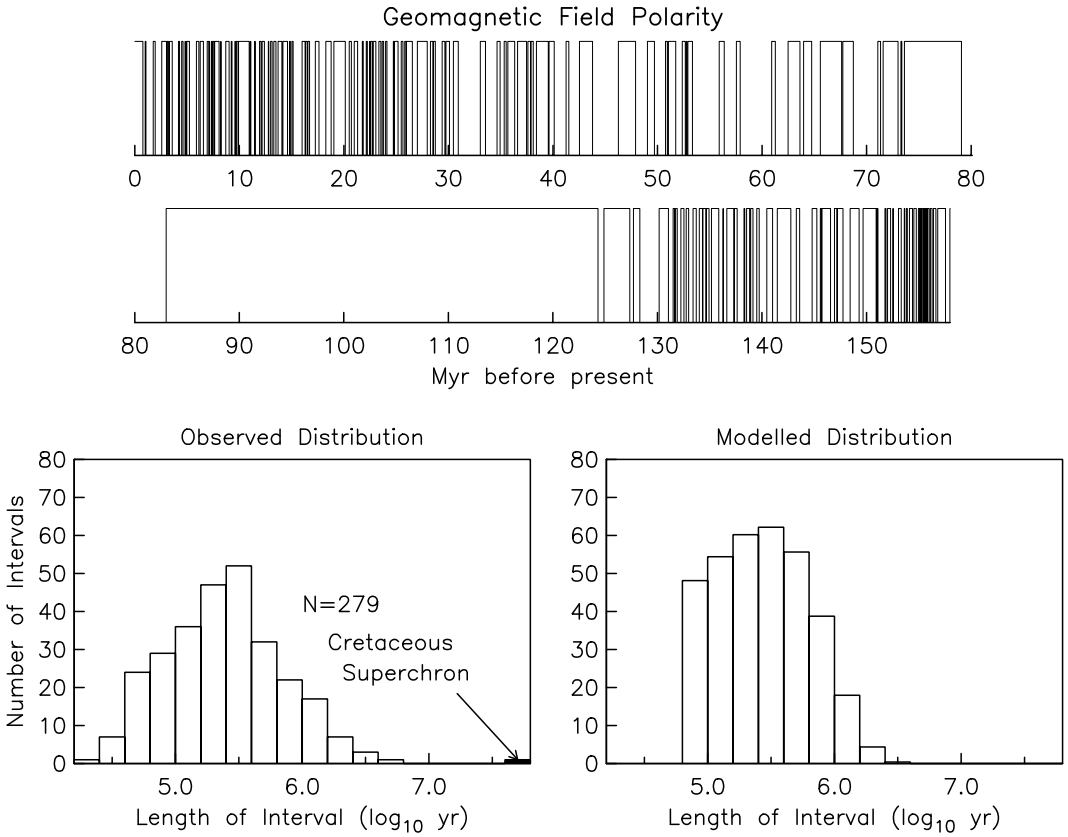

Figure 1.4

In statistics questions of this type are called **hypothesis testing**, because we can only say that a behavior is unusual in comparison to some particular hypothesis. In statistics, the meaning of "hypothesis" is more limited than in general usage: it refers to a particular probability model. Because we are dealing with probability, we can never say that something absolutely could not have occurred, only that it is very improbable. The mode of reasoning used for such tests is somewhat difficult to get used to because it seems backwards from the way we usually reason: rather than draw a positive conclusion, we look for a negative one, by first creating a model that is the opposite of what we want to show. This is called the **null hypothesis**; we then show that the data in fact make it very unlikely that this hypothesis is true.

For this example, a null hypothesis might be the probability model that there is an equal chance of a reversal in any 40,000 year interval—we choose this time because shorter reversals do not seem to be common, either because they do not happen or because they are not well recorded in marine magnetic anomalies. We call this chance (or probability) p, and we assume it is the same over all times: these assumptions combine to form our hypothesis, or (again) what we could call a stochastic model. If we estimate this p using the distribution of intervals, we find that it is (roughly) 0.1. This gives the histogram of inter-reversal times shown at the lower right of Figure 1.4. This is actually not a very good fit to the data but we can take it as a first approximation. Now, 40 Myr has 1000 intervals of 40 kyr; the probability of there not being a reversal over this many intervals consecutively is $(1-p)^{1000}$, or 2×10^{-46}. So, over the 160 Myr of data we have (four 40-Myr intervals), we would expect to get such a long reversal-less span about one time out of 10^{45}—which we may regard as so unlikely that we can reject the idea that p does not change over time. Of course, we only have the one example, so it is (again) arguable whether or not saying "one time out of N" is the right way to phrase things; but given how small the probability is in this case, we may feel justified in any case in rejecting the hypothesis we started out with. But, we should always remember that our decision to reject depends on a judgment about how small a probability we are willing to tolerate—and this judgment is, in the end, arbitrary.[3]

1.4. Distance Measurement (II): Error Bounds

Our discussion in Section 1>2 of the stochastic model for the distance measurement was about the "best estimate" of the parameter m. But this is not the only question we could ask relating to m; we might reasonably also ask how well we think we know it—that is, in the conventional phrasing, how large the error of $\hat{m}$ is. This question actually is itself a hypothesis test, or rather a whole series of such tests, for each of which the hypothesis is "m is really equal to the value ____; given our model, is this compatible (that is, likely) given the data observed?" If the assumed value were 50320 mm, or 50330, the answer would be, not very likely; if the assumed value were 50326.8, the answer would be, quite possible. We can in fact work out what this series of tests would give us for any assumed value of m; then we choose a probability value corresponding to "not very likely", and say that any value of m that gives a higher value from the hypothesis test is acceptable. This gives us, not just a value for m, but what is usually more valuable, a range for it, this range being between the *confidence limits*.

1.5. Predicting Earthquakes: A Model Misapplied

We close with an example of misapplied statistics leading to a false conclusion—and an expensively false one at that. Our example is, fittingly, one of earthquake prediction, a field that has more examples of inept statistical reasoning (as well as blissful unawareness of the need for such reasoning), than any other branch

[3] A more thorough analysis (C. Constable, On rates of occurrence of geomagnetic reversals, *Phys. Earth Planet. Inter.*, **118**, 181-193 (2000)) shows that the Poisson model can be used, provided we make the probability time-dependent. A model consistent with the data has this probability diminishing to a very low value during the Cretaceous Superchron, and then increasing to the present.

of geophysics.

The particular case was that of the repeating earthquakes at Parkfield, a very small settlement on the San Andreas fault in Central California. Earthquakes happened there in 1901, 1922, 1934 and 1966; seismometer records showed the last three shocks to have been very similar. Nineteenth-century reports of felt shaking suggested earthquakes at Parkfield in 1857 and 1881. This sequence of dates could be taken to imply a somewhat regular repetition of events.

You might think, from what we discussed in the previous section, that the relevant data would be the times between earthquakes—and you would be right. However, the actual analysis took a different approach, shown in the left panel of Figure 1.5: the event numbers were plotted against the date of the earthquake, and a straight line fit to these points. The figure shows two fits, one including the 1934 event and the other omitting it as anomalous. If the 1934 event is included, the straight line reaches event number 7 in 1983 (the left-hand vertical line); this was known not to have happened when the analysis was done in 1984. Excluding the 1934 event yielded a predicted time for event 7 of 1988.1, ±4.5 years.

Partly because of this prediction, which seemed to promise a payoff in the near future, a massive monitoring effort was set up around Parkfield, which was continued long after the "end" of the prediction. The earthquake eventually happened in September 2004, 19 years "late".

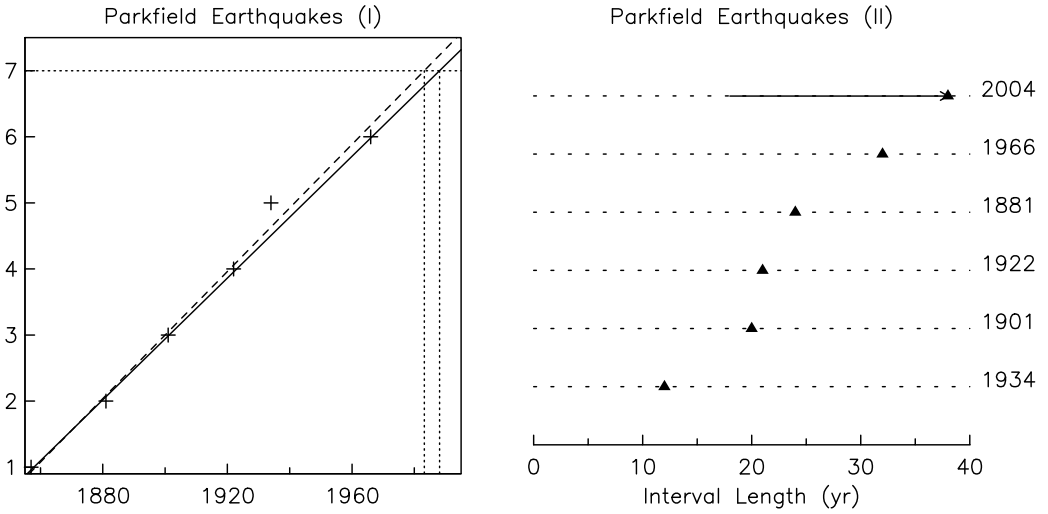

Figure 1.5

What went wrong? The biggest mistake, not uncommon, was to adopt a set of "standard" methods without checking to see if the assumptions behind them were appropriate: an approach often, and justly, derided, as the cookbook method. In this case, the line fitting, and the range for the predicted date, assumed that both the x- and y-coordinates of the plotted points were random variables with a probability distribution somewhat similar to equation (1). But the event numbers, 1 through 6, are as nonrandom as any sequence of numbers can be; the dates are not random either, for they have to increase with event number. A more careful analysis of the series shows that while the time predicted for the next earthquake in 1984 would still be

close to 1990, the range of "not-improbable" times would be much larger—the earthquake was nowhere near as imminent as the incorrect analysis suggested. No doubt some level of monitoring would have been undertaken anyway but a more thoughtful approach might have been taken if the statistical analysis had been done properly.[4]

A simple way of seeing what a better analysis would give is suggested by the right-hand plot in Figure 1.5, which shows the inter-event times ordered by size, and labeled by the date when each ended. The longest interval has an arrow extending from 1984 (when the prediction was made) to the actual time of the earthquake. It seems clear that the range of prior inter-event times is such that the most reasonable prediction for the next earthquake in 1984 would have been would have been "probably soon, but a 10-20 year wait might well be expected".

The lesson is that you should learn not just a set of techniques to use, but also when *not* to use them. In this case, one approach, and one kind of plot, were seriously misleading; a different presentation of the data would immediately have led to different conclusions.

[4] The original prediction was by W. Bakun and T. V. McEvilly, Recurrence models and Parkfield, California, earthquakes, *J. Geophys. Res.*, **89**, 3051-3058 (1984). The correct analysis is Y. Y. Kagan, Statistical aspects of Parkfield earthquake sequence and Parkfield prediction experiment, *Tectonophys.*, **270**, 207-219 (1997). This particular error is not unique to the Parkfield analysis; see Stephen M. Stigler, Terrestrial mass extinctions and galactic plane crossings, *Nature*, **313**, 159 (1985).