CHAPTER 1

PROBABILITY, STATISTICS, AND REALITY

> If your experiment needs statistics, you ought to have done a better experiment.
>
> ERNEST RUTHERFORD, attributed

> 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 worshiping very strange gods.
>
> JOHN ZIMAN, *Public Knowledge: an Essay Concerning the Social Dimension of Science* (1968)

1.1 The Aim of the Exercise

The quotations above give some of the rationales for a course devoted to statistical methods for geophysics. First, for many geophysical topics we can only observe, not do experiments: for example, we have to understand the magnetic field of the Earth as it is, not as it might be if we changed part of the system that produces it. So Rutherford's dictum is unhelpful. Statistical methods are certainly not confined to observational science – even nuclear physicists use them[1] – but they are central to working with geophysical data, and learning them is part of learning to be a geophysicist.

Our second quotation, and the epigraph for the text, illustrate that these methods are difficult to learn and to do correctly: hence the need for a course, and this text. Much of the difficulty comes from the challenge of thinking correctly about random events: a challenge that benefits casinos

[1] And have for as long as there has been nuclear physics: for example, using the statistics of a Poisson process (Section 3.4) to show the randomness of radioactive decay **?**.

and frustrates students. We hope this course will help you with learning about statistics and how to use it, 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 cover only a few topics; we have selected ones that cover much of what is needed in geophysics. The analysis of data arranged in time or space deserves fuller treatment, so much of that is postponed to the second part of the course:

We neither seek nor shirk mathematics, though we rarely aim for rigor, and mostly do not include proofs of results. An adequate mathematical background would be familiarity with calculus, at least through multivariate calculus, and with vectors, vector spaces, and matrices. We aim to provide an understanding of concepts and terminology that should be useful if you need to learn more.

In the rest of this chapter we try to give the flavor of statistical reasoning by some examples – including one case of it being applied badly. These examples introduce terms from probability and statistics: some, we hope, will be familiar, though others may not be. We use **boldface** when we introduce these, and again when we re-introduce and define them in later chapters.

1.1.1 What Kind of Problem Are We Trying to Solve?

Our first two examples are datasets that illustrate different settings in which statistical reasoning would be appropriate. The first dataset is two years of measurements between a pair of geodetic markers about 50 meters apart; in the left-hand panel in Figure 1.1. We summarize these using a **histogram** to show how many data are in various intervals. From this plot we see that the data clump around a value of 50327 mm, and are spread around this over a range of ± 1 mm. The usual way to understand this spread we call the "physics-lab" view: there is a single, true value of the distance, and we get numbers scattered around this only because of "errors" in the measurements. Statistics then becomes just a tool for dealing with something (errors) that just gets in the way of the truth.

Sometimes this view is useful; but more often it is not. The right-hand panel of Figure 1.1, is another histogram, this one of the lengths of segments of oceanic spreading centers, a segment being terminated by some other kind of plate boundary. Here there is no "true" length, and the idea

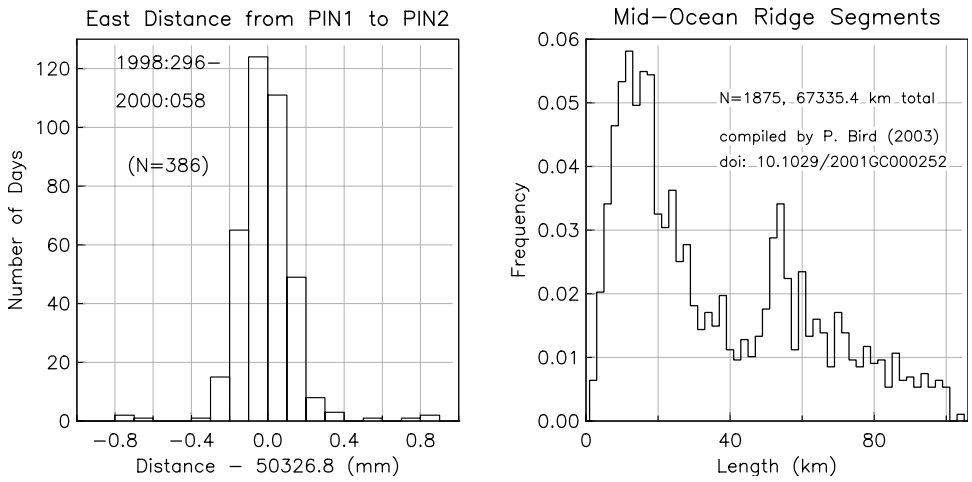

Figure 1.1: Left: histogram of estimated distance between two continuous GPS stations (PIN1 and PIN2), found using L1/L2 processing with no estimation of atmospheric zenith delays. See **?** for more information. Right: histogram of lengths of oceanic spreading centers, from the digitized plate boundary of **?**.

of measurement error is simply inappropriate.[2] Rather, how the numbers are distributed is itself information about spreading centers.

What these examples share is numbers that are somehow scattered or varying. We use probability theory, and the statistical methods that flow from it, because this is a mathematical construct well-suited to dealing with such variations. Our purpose in applying this theory, is to develop a mathematical model for a dataset. Such models are called **probability models** or **stochastic models**.

An Example of a Stochastic Model

To show what we mean by a probability model, we start with a made-up "toy model". Table 1.1.1 gives 100 pairs of numbers (x_1, x_2); these numbers vary considerably, though (by design) they are all integers. The first thing you should do with any dataset is to plot it: it is foolish not to take advantage

[2] There is a kind of error possible because we do not have complete maps of all spreading centers, and the better-surveyed a spreading center is, the more likely some break in it will be found – long segments may just be ones that do not have many ship tracks crossing them; without the detail available from echo sounding, small offsets in spreading centers are not otherwise detectable.

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

Table 1.1:

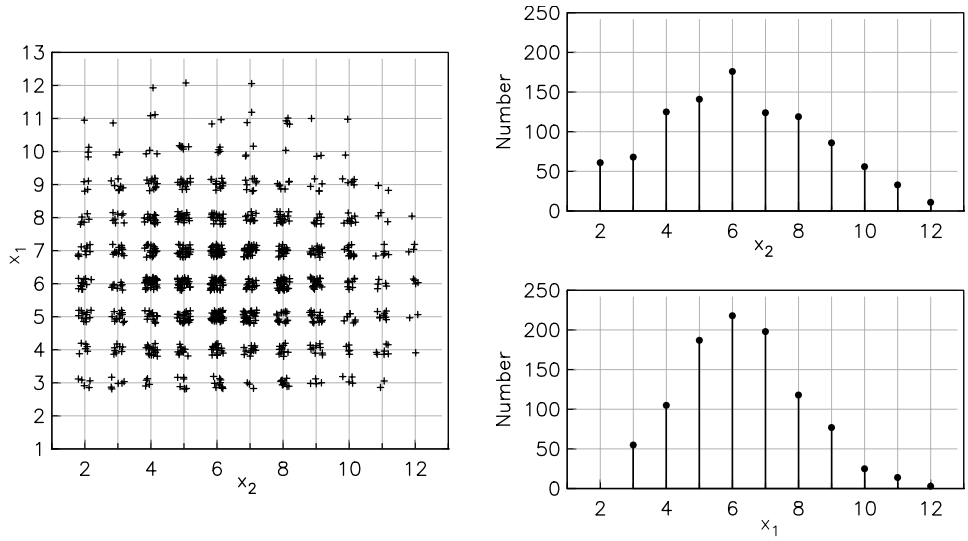

Figure 1.2: Left: plot of 1000 data pairs produced by an invented probability model. The actual values are all integers, so we have moved each point away from its value by a small random amount: a plotting trick known as **dithering**, which allows us to see the relative number at each value much better. Right: the number of points with particular values of x_1 and x_2.

of how much better humans are at understanding visual depictions than numerical lists. Figure 1.2 is a plot of more data pairs from our toy model, and shows a clear pattern to the relative occurrences of the different values. The model used to produce these data can be described by the following rule: 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; then take two numbers, each with an equal chance of being 1, 2, 3, 4, 5, or 6, and add these up to get the value of x_2. We could produce the numbers by rolling tetrahedral and cubical dice; the values listed and plotted actually came from a software simulation of that procedure.

Our brief description is an example of how a **probability model** can explain data. More generally, a probability model is a mathematical construction that we develop to explain some set of observations. Such a model includes some aspects that are "random", meaning that they do not have definite values, but only different chances of taking on different values. In our toy model we started with elements that have equal chances of taking on integer values within certain ranges, and combined them to get our final result. These elements are called **random variables**; much of developing probability models comes from deciding what parts of your data are best modeled by this type of "variable variable", and which by variables that 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".

1.1.2 Distance Measurement (I): Point Estimation

For our first illustration of a statistical procedure, we return to the GPS data set of Figure 1.1, the histogram of which we repeat in the left panel of Figure 1.3. Drawing a more precise conclusion than that the data are "clumped" requires us to assume that the data can be represented by a mathematical model that includes random variables. This is a very big step – and, like most applications of mathematical models to the real world, not one that can be justified in simple terms. 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 particular probability model might be appropriate – or it might not.

The specific mathematical model we adopt is to regard the length as a

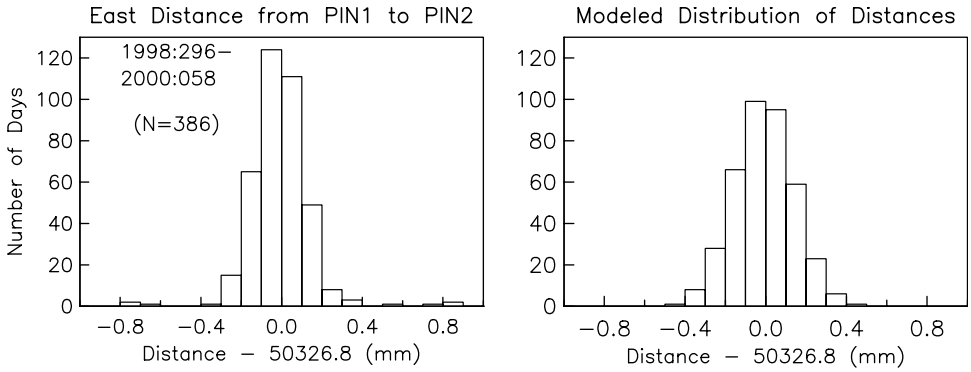

Figure 1.3: Left: histogram of distance data, repeated from Figure 1.1. Right: histogram expected for a stochastic model that has been fit to these data.

random variable X; this variable is described by the equation

$$\Pr[X < x] = \frac{1}{\sigma(2\pi)^{1/2}} \int_{-\infty}^{x} \exp\left[-\tfrac{1}{2}\left(\frac{u-m}{\sigma}\right)^2\right] du \tag{1.1}$$

The right-hand side of this equation should pose no difficulty: it describes a function of x that depends on two parameters m and σ, both of which have definite (but at this point unknown) values. The novelty, both in notation and in concept, is on the left-hand side. The expression there 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. In applying it to this particular dataset, we can most usefully view it as a way of representing the fraction of times that we get a particular result if we repeat something, which is called the **frequency of occurrence**. The simplest example of this interpretation is when we say that the probability of getting heads on tossing a fair coin is 0.5.[3] Certainly this **frequency interpretation** is the most straightforward way to connect equation (1.1) to the data summarized in Figure 1.3.

So, we have data, and we have assumed a 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

[3] Note that we give probabilities as values from zero to one, *not* as percentages. You should too.

rules inspired by real-world examples. Statistics, while it makes use of probability theory, is something else, namely the application of this theory to data so as to draw actual conclusions. Though for historical reasons nobody calls Statistics a branch of applied mathematics, that is, we think, the best way to classify it.

Given our model, as expressed by equation (1.1), and given these data, an immediate statistical question is, what are the "best" values for m and σ? Assuming that (1.1) in fact correctly models these data, the formulas for finding the "best" m and σ, (symbolized by $\hat{m}$ and $\hat{\sigma}$) are

$$\hat{m} = \frac{1}{n}\sum_{i=1}^{n} d_i \qquad \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \hat{m})^2 \tag{1.2}$$

where $d_1, d_2, \ldots d_n$ are the n data. Applying these expressions to the data from which Figure 1.2 was drawn, we get $\hat{m} = 50326.795$ and $\hat{\sigma} = 0.175$. The parameters m and σ in equation (1.1) are called the **mean** and **standard deviation**; the values $\hat{m}$ and $\hat{\sigma}$ that we get from the data using the formulas in (1.2) are called **estimates** of these.[4] In Chapter 5 we will describe in what sense these estimates might be termed "the best;" finding and evaluating procedures such as equation (1.2) is the area of statistics known, unsurprisingly, as **estimation theory**. Finding the best values of parameters such as m and σ is called **point estimation**: perhaps the commonest, but not always the most relevant, statistical question to ask and answer.

So far we have, intentionally, said nothing about "errors"; but if we interpret these data representing a true value contaminated by errors, we get a formally similar but philosophically quite different statement, which is that the data can be modeled as

$$d_i = t + e_i \tag{1.3}$$

where t is the true value of the distance, and the e_i are the errors. We then can model the errors as being represented by a random variable E, the probability model for which is

$$\Pr[E < e] = \frac{1}{\sigma(2\pi)^{1/2}} \int_{-\infty}^{e} \exp\left[-\tfrac{1}{2}\left(\frac{u}{\sigma}\right)^2\right] du \tag{1.4}$$

Mathematically, equations (1.3) and (1.4) are equivalent to equation (1.1) if we have $t = m$, which means that if we assume these two equations,

[4] Using a superposed hat to denote an estimate of a variable is standard in statistics.

our "best estimate" of t is $\hat{m}$. This is, by far, the most frequent statistical computation scientists make: 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; in mathematical terms t is some number, not (like the e's) a random kind of thing.

We will avoid this approach, which (as we have noted) is often inappropriate in geophysics; instead we will pose problems in terms of equations such as (1.1). In such expressions the non-random variables are all on the same footing: they are 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 (1.1); the right side of Figure 1.2 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. These are quite valid objections: the data are not completely represented by this model. We could find a better model; but if our goal is the get a best estimate of the length, we should instead consider what is known as **robust estimation**: point estimation procedures that are not affected by small errors in the model.

1.2 Magnetic Reversals, Earthquakes, and Hypothesis Testing

In the previous section we estimated parameters of a statistical model; but how might we decide the underlying question of whether or not our model is valid for the data? This is a different kind of statistical question. For our first example, we use the data plotted in Figure 1.4: reversals of the Earth's magnetic field for the interval 0–159 Ma. This is an example of a particular kind of geophysical data that happens over time, this kind being called a **point process**. This name is used because the behavior over time is defined by the particular times, or points, at which something happens: for example, a field reversal, an earthquake, or a disk crash.

The simplest probability model for this behavior is called a **Poisson process**. An approximate description is easy: we divide time into equal

intervals, and assume the probability of an event in each interval is the same value p; in particular, p is the same whether it it has been a long time or a short time since the previous event. This specification says nothing about absolute time, so it cannot describe the times of the reversals; what it instead describes is the lengths of the intervals between them. The lower left histogram in Figure 1.4 is of the intervals observed for the magnetic reversals, plotted on a logarithmic scale because the range is from 2×10^4 to 4×10^7 years. The longest interval (the "Cretaceous Normal Superchron" or CNS) ran from 124 Myr ago to 83 Myr ago. Looking at the time series, and even more the histogram, raises an obvious question: is this very long interval "unusual" compared to the others? If we decide that it is, we could argue more strongly that the core dynamo (the source of the field) changed its behavior during the CNS.

In statistics this kind of question about "unusualness" is an example of a **hypothesis test** – the reason being that we say that a behavior is unusual only if we compare it against what would be expected from some probability model, which we call a "hypothesis". The random element that is basic to any probability model means that we cannot say that some data make a model impossible, only that it is very improbable that the data could come from a specific model. Testing for improbability involves reasoning in an unfamiliar way, one that can seem backwards. We first create a model (called the **null hypothesis**) opposite from the one that would describe what we want to show. We then look for a negative conclusion, namely that the observed data are very unlikely to come from such a null model – so the (other) model we want to show is more likely to be a valid description.

For testing if the CNS interval is unusual compared to the others, we chose a null hypothesis (probability model) that describes all of these others. A simple model would be the kind of Poisson-like process we described above, with an equal probability p of a reversal in any 40,000 year interval; we choose this basic length because shorter intervals are rare. An estimate of p from the data is $p = 0.1$; this would produce the histogram of interval lengths shown in the lower right of Figure 1.4; at least a crude approximation to the data. The 40 Myr of the CNS 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}. This is so very unlikely that we can be comfortable rejecting the Poisson model for the CNS period, even though, given that there is only one field, it may seem odd to say "one time out of

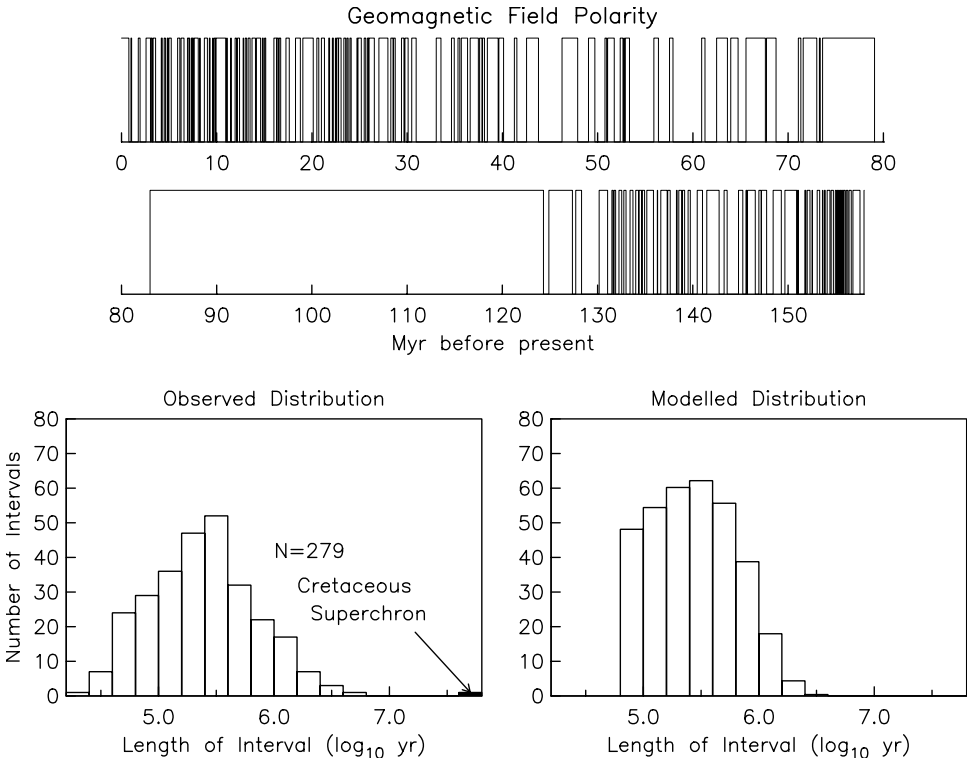

Figure 1.4: The top two plots show the reversals of the Earth's magnetic field, mostly reconstructed from marine magnetic anomalies, over the last 159 Myr; a positive value means that the dipole has the same direction as now, and negative ones that it was reversed. The lower plots are histograms of the intervals between reversals, binned in log time: on the left, the histogram for the data, and on the right, one for for a Poisson process with the same rate as the data, and a minimum interval of 40,000 yr.

N". We should always remember that in using a low probability to reject a model we are making an arbitrary judgment about how small a probability we will tolerate.[5]

We use another data set, earthquakes in California, to introduce another example of hypothesis testing, this time to rule out the "non-null" hypothesis. It is sometimes said that large California earthquakes tend to occur in the early morning because this was true for shocks in 1906 (San Francisco), 1971 (San Fernando), 1992 (Landers), 1994 (Northridge), and 1999 (Hector Mine). Early morning is a good time to have an earthquake, since most people were at home, which in California is a relatively safe place to be. We guessed this behavior from a few (selected) examples); we can ask if it is supported by a full earthquake catalog.

Figure 1.5 shows the data, which we plot in four different ways to emphasize that there are often more effective ways to plot data than what might seem most obvious. The obvious plot as in Figure 1.1 is a histogram (upper left plot in Figure 1.5), but this places the two end times far apart, when they are actually close. One cure for this is a radial histogram, or **coxcomb plot**, shown in the upper right plot in Figure 1.5; the sizes are scaled so that the number of events is proportional to the area of each sector.

Both of these plots have a defect (shared with Figure 1.1: we have to decide how big to make the bins. Too small, and there will be wide variations in the sizes, too large, and we might lose interesting detail. We can say, with emphasis, *never bin data unless they come in bins or you have no choice*. Fortunately, for the earthquake data we do have a choice, which we show in the bottom left plot in Figure 1.5: use the magnitude (certainly important) as a second variable and plot the individual earthquakes as a function of magnitude and of time of day. We repeat part of the left-hand side on the right so we can see the distribution around the day boundary. But we can do better if we use a polar plot, where time is the angle, and magnitude is related to the radius. If we make the radius larger for smaller magnitudes, the many smaller earthquakes are spread over a larger area and so are easier to see. In the last two plots we use different symbols to separate earthquakes recorded on seismometers from those known only from felt reports: this plot easily combines three types of information.

[5] A more thorough analysis by **?** shows that a Poisson model can be used, provided the probability p can change with time: p has a very low value during the Cretaceous Superchron, and increases from then to the present.

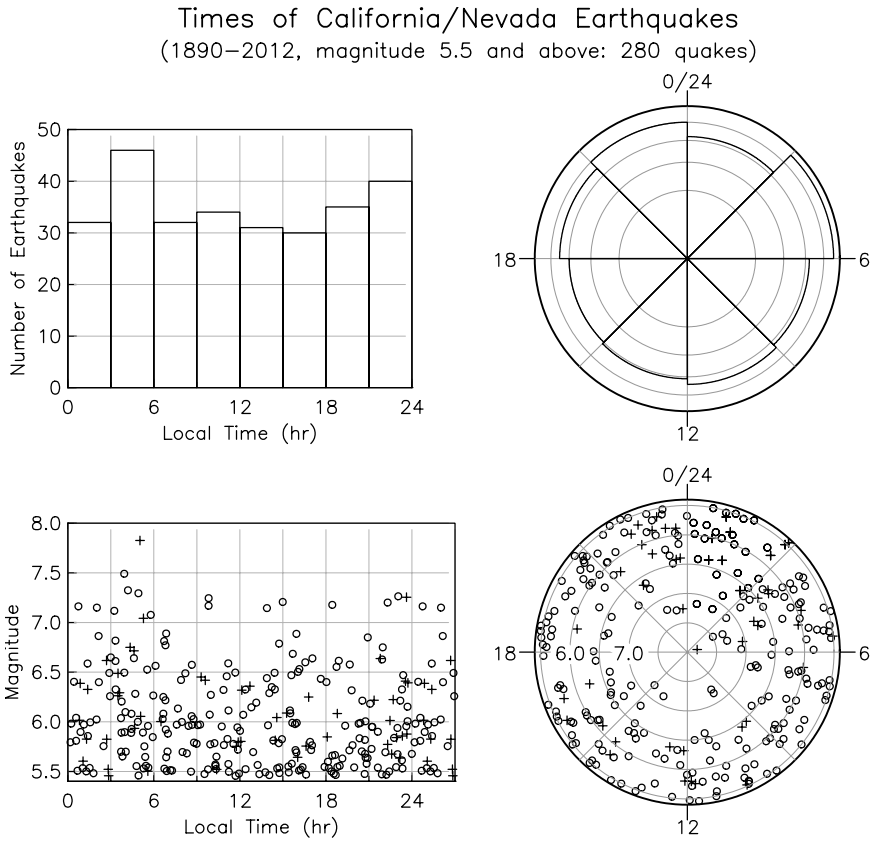

Figure 1.5: Four plots of the local-time distribution of earthquakes from 1890 on in California and Nevada, with magnitude 5.5 or larger, The data are from the catalog of the Working Group on California Earthquake Probabilities, combined with the list in **?**, and with immediate aftershocks removed. In the two bottom plots, pluses are for the 41 shocks during 1890-1909, circles for the 239 during 1910-2011.

Certainly this last plot does not suggest any concentration around one range of times. We postpone our actual test to Chapter 6.

1.3 Distance Measurement (II): Error Bounds

In Section 1.1.2 we used a stochastic model for the distance measurement to make a "best estimate" of the parameter m. But usually we do not want only this, but also some characterization of how well we think we know it – in the conventional phrasing, how large the error of $\hat{m}$ is. This question can actually be framed as a whole series of such hypothesis tests, for each of which the hypothesis is "m really is equal to a particular value". Given a probability model we can ask if this is likely given the data observed. In this case, if the value was assumed to be 50320 or 50330 mm, 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 (low) probability value corresponding to "not very likely", and say that we will accept any value of m for which the hypothesis test gives a higher value. We thus get not just a value for m, but (more usefully) a possible range for it, known as the **confidence interval**. As with the point estimate, our result will depend on the stochastic model we choose; if the model is bad, our conclusions will be too.

1.4 Predicting Earthquakes: A Model Misapplied

We close with an example where bad conclusions were drawn from an incorrect model – and these conclusions were not just false, but costly. Our example comes from the field of earthquake prediction, which probably has more examples of inept statistical reasoning, and of blissful unawareness of the need for such reasoning, than any other branch of geophysics.

This prediction was for an earthquake 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

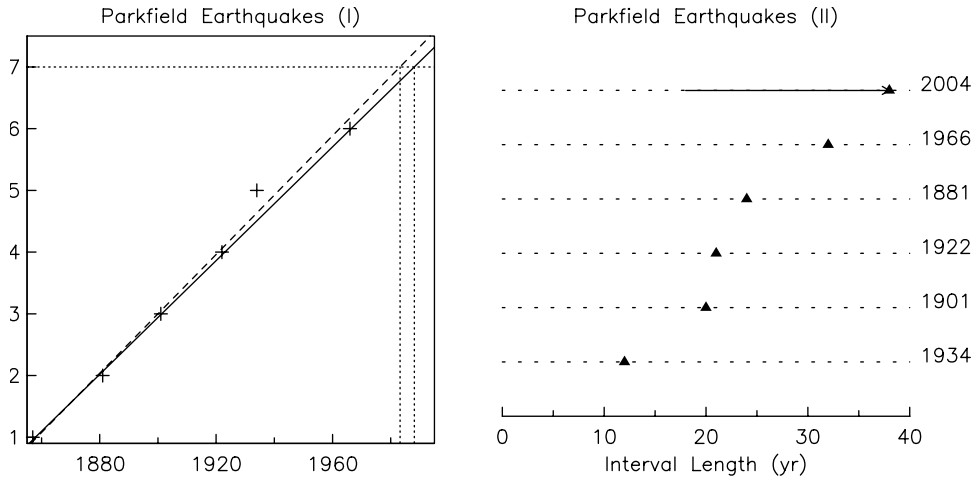

Figure 1.6: Left: times of Parkfield earthquake, 1857 through 1966, plotted against earthquake number, with lines fit to the data values with the 1934 earthquakes included and excluded, and the resulting times predicted for the next earthquake. Right: intervals between earthquakes (starting with the 1881-1901 interval), ranked by interval length. The solid line for the 2004 earthquake shows the time between the original prediction and the actual event.

felt shaking suggested earthquakes at Parkfield in 1857 and 1881. This sequence of dates could be taken to imply a moderately regular repetition of events.

You might think, from our discussion in Section 1.2, that a model should be derived for the times between earthquakes – and you would be right. But the actual analysis took a different approach, shown in the left panel of Figure 1.6: 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 seven 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 seven of 1988.1, with an "error" of 4.5 years.

Partly because this prediction promised a payoff in the near future, a large monitoring effort was set up around Parkfield. This effort continued long after the "end" of the prediction – very long after, because the seventh earthquake happened in September 2004, 19 years "late".

What went wrong? The biggest, and all too common, mistake was to

adopt "standard" methods without checking to see if the assumptions behind them were appropriate: an approach often, and justly, derided as the "cookbook method". The statistical methods used to fit the line and find 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 (1.1). But the event numbers, one through six, are as nonrandom as any sequence of numbers can be; and the dates are not random either, since they increase with event number. A more careful analysis of the series showed that, in 1984, a time near 1990 would be the best point estimate of when the next earthquake would happen. But the range of "not-improbable" times would be much larger than nine years, so the next 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.[6]

But if the data had been plotted properly this error would most likely have been avoided. The right-hand panel in Figure 1.6, is an alternative plot: it shows the intervals, 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. The range of prior intervals suggests that in 1984 a reasonable prediction for the next earthquake would have been would have been "probably soon, but a 10-20 year wait might well be expected".

This example shows that you need to learn not only techniques to use, but also when *not* to use them; it also shows that how you plot data can easily affect your conclusions.

[6] The original prediction was by **?**; the correct analysis is **?**. This particular misapplication of statistics was not unique to the Parkfield analysis; see **?**.