CHAPTER 2

PROBABILITY AND RANDOM VARIABLES

> In statistics it is a mark of immaturity to argue overmuch about the fundamentals of probability theory.
>
> M. G. KENDALL and A. STUART *The Advanced Theory of Statistics* (1977)

2.1 Introduction

In this chapter we introduce a few concepts from probability theory,[1] starting with the basic axioms and the idea of conditional probability. We next describe the most important entity of probability theory: the random variable; this requires that we introduce the probability density function (and the related distribution function) needed to describe it. We then define means, variances, expectations, and moments of these functions (and so of the random variables they describe), and show how to do arithmetic with random variables. We close by applying some of these developments to demonstrating the Central Limit Theorem, which provides the main justification for using the Normal probability density function (equation 1.1) to model data.

[1] We use the term **probability theory** for a branch of mathematics; this is the general usage. Kendall and Stuart call this the calculus of probabilities, which allows them to make the useful distinction between this bit of mathematics, and what they call probability theory, which is how this mathematics applies (or not) to the real world – something we discuss in Section 2.2.

2.2 What is Probability?

There is a long-running dispute over what, in the real world, the formal mathematics of probability theory correspond to. The two commonest views are:

- The **frequentist** interpretation, which is that the probability of something corresponds to what fraction of the time it happens "at random" or "in the long run". This might well be called the casino interpretation of probability, since that is one place where it makes sense. But there are many others in which it does not; as we noted in Chapter 1 geophysics provides many examples. It might make sense to talk about the probability that the next earthquake in California will be bigger than some amount, since there are lots of earthquakes; but it is much less clear how to apply frequentist concepts to (say) the Earth's gravitational field: there is only one.

- The **Bayesian** or **subjective** interpretation, in which the probability of something corresponds to how likely we think it is: probabilities represent states of mind. This view has led to a distinctive set of methods for analyzing data.

We have already hinted at our preference, which is that it is actually meaningless to ask what probability "really is". Rather, it seems most useful to regard probability as a mathematical system; like other parts of mathematics it can be used to model certain aspects of the real world. In this view, probability is no different from other mathematical idealizations that are used for models: for example, in studying seismic waves, we represent the Earth by an elastic solid – which is just as much a mathematical idealization. If we simply take probability as some mathematics used in a model, there is no problem in having it represent more than one kind of thing, so both Bayesian and frequentist interpretations can each be valid or invalid, depending on what it is we choose to model.

2.3 Basic Axioms

So, what is the mathematics of probability? The basic concepts and axioms were developed by Kolmogorov using set theory; though the names used are meant to suggest the "casino" model, the mathematics does not require

this. We start with the idea of a **sample space** Ω: a set whose elements are subsets containing all possible outcomes of whatever it is we are proposing to assign probabilities to. Examples of outcomes are heads or tails, a value from a throw of dice, normal or reversed magnetic fields, or the results of doing some experiment or making some observation. Note that outcomes need not be numbers.

We denote each set of outcomes by a letter (e.g., A), and the **probability** of that set of outcomes by $\Pr[A]$. Then the rules for probabilities are:

- $\Pr[\Omega] = 1$; the probability of all the outcomes combined is one, indicating that some outcome has to happen (true since Ω includes all outcomes by definition).
- $\Pr[A] \geq 0$; probabilities are positive.
- If two sets of outcomes are disjoint (mutually exclusive) then $\Pr[A_i \cup A_j] = \Pr[A_i] + \Pr[A_j]$: the probability of the combination (union of the sets) is the sum of the individual probabilities.[2] That is, if outcome A_i precludes outcome A_j and vice-versa, the probability of having either one is the sum of the probabilities for each (think of throwing a die, which has six disjoint outcomes).

All of these rules are pretty good fits to the kinds of things we are attempting to model; they are almost intuitive to how we think about randomness. But these few axioms are enough to produce the whole theory.

2.4 Conditional Probability

Things become slightly more interesting (because less obvious) when we introduce the concept of **conditional probability**. This is written as $\Pr[A|B]$, meaning "The probability of outcome set A given that we have outcome set B", the last part of which is sometimes phrased as "given that outcome B is true". The relation for this is that $\Pr[A|B]\Pr[B] = \Pr[A \cap B]$: the probability that A and B are both true is the probability of B being true, times the probability of A being true given B. This is more usually written so as to define conditional probability:

$$\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]} \tag{2.1}$$

[2] Remember that $A \cup B$ is the union of A and B; $A \cap B$ is the intersection of A and B.

Conditional probability leads to the concept of two sets being **independent**: A and B are independent if $\Pr[A|B] = \Pr[A]$, which is to say that the probability of A does not depend on whether B has happened or not.[3] This means, from 2.1, that $\Pr[A \cap B] = \Pr[A]\Pr[B]$: if A and B are independent, the probability of having both A and B is the product of their individual probabilities. This rule is easily abused, since it is all too tempting to decide that events are independent when they actually are not.

2.4.1 Applying Conditional Probabilities: Was That a Foreshock?

We can apply conditional probabilities to a geophysical problem of actual social significance by asking what we should do if a small earthquake occurs close to (say) the San Andreas fault, given that it might be either a foreshock to a major earthquake on this fault, or just be a small shock that happened there by chance (a "background" earthquake). The full treatment [**?**] becomes rather complicated, but a simplified version illustrates the procedure, and the use of conditionals. Our set of possible outcomes is three events:

- A background earthquake has occurred: (B).
- A foreshock has occurred: (F).
- A large (so-called characteristic) earthquake will occur: (C).

If a small background shock were to coincidentally happen just before the characteristic earthquake, we would certainly class it as a foreshock. So, B and C are disjoint: they cannot both occur. The same holds true for B and F: we can have a foreshock or a background earthquake, but not both.

The probability that we want is the conditional probability of C, given either F or B (because we do not know which has occurred). This is, from equation (2.1),

$$\Pr[C|F \cup B] = \frac{\Pr[C \cap (F \cup B]]}{\Pr[F \cup B]} \tag{2.2}$$

Because F and B are disjoint, the probability of their union is the sum of the individual probabilities (axiom 3), allowing us to write the numerator

[3] Often this is called **statistical independence**; the extra adjective is confusing, since there is no use of statistics in the definition.

as

$$\Pr[(C \cap F) \cup (C \cap B)] = \Pr[C \cap F] + \Pr[C \cap B] = \Pr[C \cap F]$$

where the disjointness of C and B eliminates the $\Pr[C \cap B]$ term. By the definition of conditional probability,

$$\Pr[C \cap F] = \Pr[F|C]\Pr[C] \tag{2.3}$$

where $\Pr[F|C]$ is the probability that a mainshock is preceded by a foreshock. Using the disjointness of F and B again, the denominator becomes

$$\Pr[F \cup B] = \Pr[F] + \Pr[B] \tag{2.4}$$

Because a foreshock cannot, by definition, occur without a mainshock, the intersection of C and F is F, and therefore

$$\Pr[F] = \Pr[F \cap C] = \Pr[F|C]\Pr[C] \tag{2.5}$$

We can use equations (2.3), (2.4), and (2.5) to rewrite equation (2.2) as

$$\Pr[C|F \cup B] = \frac{\Pr[F]}{\Pr[F] + \Pr[B]} = \frac{\Pr[C]\Pr[F|C]}{\Pr[F|C]\Pr[C] + Pr[B]} \tag{2.6}$$

For $\Pr[B] \gg \Pr[F|C]\Pr[C]$ this expression is small (the candidate event is probably a background earthquake), while for $\Pr[B] = 0$, the expression is equal to one: any candidate earthquake must be a foreshock.

The second form of expression in 2.6 is a function of three quantities, which in practice we obtain in very different ways. $\Pr[B]$, the probability of a background earthquake, comes from seismicity catalogs for the fault zone. $\Pr[C]$, the probability of a characteristic earthquake, comes from the past history of large earthquakes on this fault, usually determined by paleoseismological studies. If we had a record of the seismicity before many large earthquakes on the fault, we could evaluate $\Pr[F|C]$ directly; but because of the limited time over which seismicity has been recoded, we do not have such a record. So we have to assume that the average of $\Pr[F|C]$ over many earthquakes on one fault is equal to the spatial average over many faults over a shorter time; perhaps not valid, but the best we can do.

2.4.2 Natural Frequencies: Another Frame for the Problem

While the algebraic manipulations in Section 2.4.1 are needed to develop the solution to the full problem, they are not the easiest way to get the

simplified result just given. Strangely, insight into problems of this sort depends very much on just how they are phrased [**?**]. For almost everyone, stating the numbers in terms of probabilities does not help intuitive reasoning; what works much better is to state them in terms of numbers of events (out of some large but very round number), an approach called **natural frequencies**.[4] We recommend this approach for explaining conditional probability methods to other people, or even to yourself (as a good way to check your algebra). To show this method for the foreshock computation, suppose we had a C every 100 years, a B 10 times a year, and half the C's had F's. Then in (say) 1000 years we would expect 10 C's, and hence 5 F's; and also 10,000 B's. So we would have 10,005 possible B's and F's, and the chance that any possible member of this class would be an F would thus be 5/10005. You can easily plug in the daily probabilities for F, C, and V into 2.6 to get the same result – which, put this way, seems almost trivial.

2.5 Bayes' Theorem

The procedures followed for the foreshock probability estimate are very similar to those used to derive **Bayes' Theorem**, a result that forms the basis for one type of statistical inference. The theorem itself is not difficult to derive. Suppose we have N *disjoint* sets of outcomes, called $B_1, B_2, \ldots B_N$, and another set A. The probability of both A and a particular one of the B's (say B_j) is, by the definition of conditional probability,

$$\Pr[A \cap B_j] = \Pr[B_j|A]\Pr[A] = \Pr[A|B_j]\Pr[B_j] \tag{2.7}$$

where you should remember that $\Pr[A \cap B_j] = \Pr[B_j \cap A]$. But, since the B's are disjoint, $\Pr[A] = \sum_j \Pr[A|B_j]\Pr[B_j]$. Combining this with 2.7, we find that

$$\Pr[B_j|A] = \frac{\Pr[A|B_j]\Pr[B_j]}{\sum_j \Pr[A|B_j]\Pr[B_j]} \tag{2.8}$$

The different parts of this expression have special names: Each B_j is called a **hypothesis**, $\Pr[B_j]$ is called the **prior probability** of B_j, and $\Pr[A|B_j]$ the **likelihood** of A given B_j.

[4] **?** tested how well this and two other approaches worked when people had to process the results (e.g. adding or multiplying probabilities); they found that expressing probabilities as "1 in [some number]" is much less easily understood – though it may be needed for very small probabilities.

All this is unproblematic; the contentiousness comes in deciding how to apply this to inference about the real world. One way to do so is to regard the Pr[B]'s as degrees of belief about a hypothesis: for example, Pr[B_1] would be our belief (expressed as a probability) that a coin is fair, and Pr[B_2] our belief that it actually has heads on both sides. Now suppose we toss the coin four times, and get heads in each case. Then A is (for this example) the result that all of four tosses give heads, the probability of which (the likelihood) is 1/16 if B_1 is true, and 1 if B_2 is true. Then equation (2.8) allows us to find Pr[$B_j|A$], the **posterior probability** of each hypothesis.

The attractiveness of this scheme is clear: we have used the data to alter our degree of belief in some fact about the world, which is what we would like to do with all data. This procedure is called **Bayesian inference**. But we have evaded one part of this: how should we set the prior probabilities? Our evasion is deliberate, since deciding on prior probabilities is complicated and controversial. So for now we put Bayes' theorem and Bayesian inference aside.

2.6 Random Variables: Density and Distribution Functions

So far we have talked about "outcomes" as things described by set theory. But most of the time, what things we want to model are described by numbers, so we introduce the idea of a **random variable**, which we denote by (say) X. It is extremely important to realize that this is *not* the same thing as the variables we know from algebra and calculus, which we will call **conventional variables**; this name is our own invention, adopted because it allows us to keep emphasizing the difference between various types of variables, and remind you that some applications call for one type, and some for the other.

A random variable is a particular kind of mathematical entity; just as vectors and scalars are different kinds of things, so random and conventional variables are not the same. Conventional variables have definite (if unknown) values, and can be described by a single number (or a group of numbers); random variables do not have any particular value, and have to be described using probabilities. We follow the convention in probability and statistics that upper case (e.g., X) denotes a random variable, while

lower case, x, denotes a quantity which always has the same value: in our terminology, a conventional variable. As is common in the statistics literature we will often abbreviate random variable as **rv**.

A common source of confusion is that these very different kinds of variables can seem to refer to nearly identical things in the world. Consider (again) rolling dice. For a particular roll of the dice, the conventional variable x describes what we actually get – this is clearly not subject to variation. But before we roll the dice, or if we merely imagine doing so, the random variable X is what we have to use to describe the outcome to be expected.

Formally, a random variable is a mapping from a sample space Ω (containing all possible outcomes) to the relevant space of numbers. For example, if Ω is the outcomes from rolling a pair of dice, the space of outcomes maps into the integers from 2 through 12. If Ω maps into some part of the real line $\Re$ (or the whole of it) X is termed a **continuous random variable**; if, as in our dicing example, the rv maps into some or all of the integers, it is called a **discrete random variable**. Either way, each element ω in Ω corresponds to a unique number $X(\omega)$.

We describe a conventional variable x with some number; how do we describe a random variable X? The answer is that we need a function, which is called the **probability density function** of X. This makes random variables much more complicated than conventional ones: a conventional variable is completely specified by a number, while to specify a random variable takes a function.

We can best understand what such a function does by looking at the most common way of plotting the relative frequency of different values of random data, the **histogram**, which we already used in Chapter 1 for four datasets. In Figure 1.1 we showed two forms of this plot: for the GPS data we plotted the number of observations in each bin; for the ridge data we divided this number by the total number of data to get relative frequencies in a normalized histogram; the values in each bin should then be about the same for different numbers of data. Such a normalized histogram is a crude version of a probability density function (**pdf**) for a random variable, which we symbolize by $\phi(x)$.

The pdf relates to probability in the following way: the probability of the random variable X lying in the interval $[x, x+\delta x]$ is given by the integral, over that interval, of the probability density. We write the probability of X

lying in this interval as:

$$\Pr[x \le X \le x+\delta x] = \text{Prob}[x \le X \le x+\delta x] = p[x \le X \le x+\delta x].$$

where we have used two other common notations for probability, Prob[], and p[]. The relationship between the pdf and the probability of X is then given by the formula

$$\Pr[x \le X \le x+\delta x] = \int_x^{x+\delta x} \phi(u)\,du$$

For any x and small interval δx this means

$$\Pr[x \le X \le x+\delta x] \approx \phi(x)\delta x + (\delta x)^2$$

so that $\phi(x)$ represents the density of probability per unit value of x in the neighborhood of x.

Probability density functions have the following properties:

- $\phi(x) \ge 0$ for all x: probabilities are always positive.
- $\int_{L_b}^{L_t} \phi(x)\,dx = 1$: X must take on some value within its permissible range. Often this range is all of the real line, with $L_b = -\infty$ and $L_t = \infty$; but sometimes it is only a part. Section 1.2 already gave an example, which is that time intervals have to be positive, so $L_b = 0$ and $L_t = \infty$. Or, if we were considering the direction of something, X has to fall within $[0, 2\pi)$.

The usual notation for a random variable X being distributed with a pdf ϕ is $X \sim \phi$. Note that for a continuous random variable X, $\Pr[X = x] = 0$ for all x: there is zero probability of X being exactly equal to any value.

If we integrate the probability density function, we get a **cumulative distribution function** (or **cdf**), which we have already denoted by $\Phi(x)$:

$$\Phi(x) = \int_{L_B}^{x} \phi(x)\,dx = \Pr[L_b \le X \le x]$$

which means that

$$\phi(x) = \frac{d\Phi(x)}{dx}$$

provided this derivative exists, and also that

$$\Pr[x \le X \le x+\delta x] = \Phi(x+\delta x) - \Phi(x) \tag{2.9}$$

Figure 2.1 shows a pdf $\phi(x)$ and its corresponding cdf $\Phi(x)$.

The cdf has the following properties:

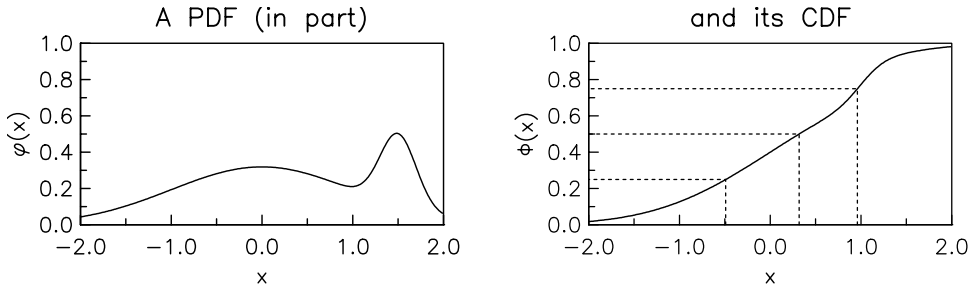

Figure 2.1: The left panel shows a possible pdf, $\phi(x)$, for a made-up distribution; the right panel shows the corresponding cdf $\Phi(x)$. The dashed lines show how to find the 0.25, 0.50, and 0.75 quantiles: these are the points on the x-axis intercepted by these lines.

- $0 \le \Phi(x) \le 1$.
- $\lim_{x\to-\infty}\Phi(x) = 0 \quad \lim_{x\to\infty}\Phi(x) = 1 \quad$ or $\quad \Phi(L_b) = 0 \quad$ and $\quad \Phi(L_T) = 1$.
- Φ is non-decreasing; $\Phi(x+h) \ge \Phi(x) \qquad$ for $h \ge 0$.
- Φ is right continuous; $\lim_{h\to0^+}\Phi(x+h) = \Phi(x)$; that is, as we approach any argument x from above, the function approaches its value at x.

While the cdf is perhaps less intuitive than the pdf, we will see that there are several advantages in using it.

The cdf also allows us to introduce what are called the **quantiles** of the distribution. Because Φ is monotonically increasing, it has an inverse Φ^{-1}. Then, if the quantile value is q, the associated value of x is

$$x(q) = \Phi^{-1}(q)$$

where Φ is the theoretical cdf. Figure 2.1 shows in graphic form the quantiles for a cdf, for q equal to 0.25, 0.50, and 0.75. The quantile for $q = \frac{1}{2}$ may be more familiar to you as the **median**; we discuss some of the others in Section 2.7.

2.6.1 The Lebesgue Decomposition Theorem and How to Avoid It

Most treatments of probability take the cumulative distribution function for X as being the more fundamental description of a random variable, using equation 2.9 for the relation to probability, and then define the pdf $\phi(x)$

as the derivative of the cdf if this derivative exists. This approach allows discrete as well as continuous random variables, through the **Lebesgue decomposition theorem**. This theorem states that any distribution function, $\Phi(x)$, can be written in the form

$$\Phi(x) = a_1\Phi_1(x) + a_2\Phi_2(x) + a_3\Phi_3(x)$$

with $a_i \gg 0$, and $a_1 + a_2 + a_3 = 1$. Φ_1 is absolutely continuous (i.e., continuous everywhere and differentiable for almost all x), Φ_2 is a step function with a countable number of jumps (that is, the sum of a finite number of Heaviside step functions, suitably scaled); Φ_3 is singular, and we ignore it as pathological. Φ_2 has the form $\Phi(x) = \sum_{x_i \le x} p_i$, where $p_i = \Pr[X = x_i]$; that is, the random variable X has a finite probability of occurring at the discrete values $x_1, x_2, x_3 \ldots$, and zero probability of having any other values. Then p_i is called the **probability mass function** or the frequency function of the random variable X; we avoid the latter term because of possible confusion with frequency in the Fourier sense. The Lebesgue theorem says that cumulative distribution functions can be used for both continuous and discrete random variables – or indeed for a combination of the two. Dice-throwing has been our standard example for a discrete rv; we could also use this kind of rv to model the probability of the number of some kind of event: for example, the number of magnitude 6 or larger earthquakes in a year which is integer-valued.

If the rv has a discrete component, the cumulative distribution function will have steps. While the derivative does not, strictly speaking, exist at these steps, we can obtain such a cumulative distribution function by integrating a pdf $\phi(x)$ that contains δ-functions, in which case ϕ is a generalized function. This approach is not usually followed in probability theory, perhaps because the standard mathematical development of that theory predates the development of generalized functions. We will be using delta-functions for many other things, so we feel free to include them in pdf's that describe a discrete rv.

2.6.2 Empirical Cumulative Distributions

We can usefully apply the cdf to show how data are distributed by plotting what is called the **empirical cumulative distribution function** or the **sample distribution function**; we denote this by $S_n(x)$. As we noted in Section 1.2, the histogram is intuitively obvious, but has a major defect: its

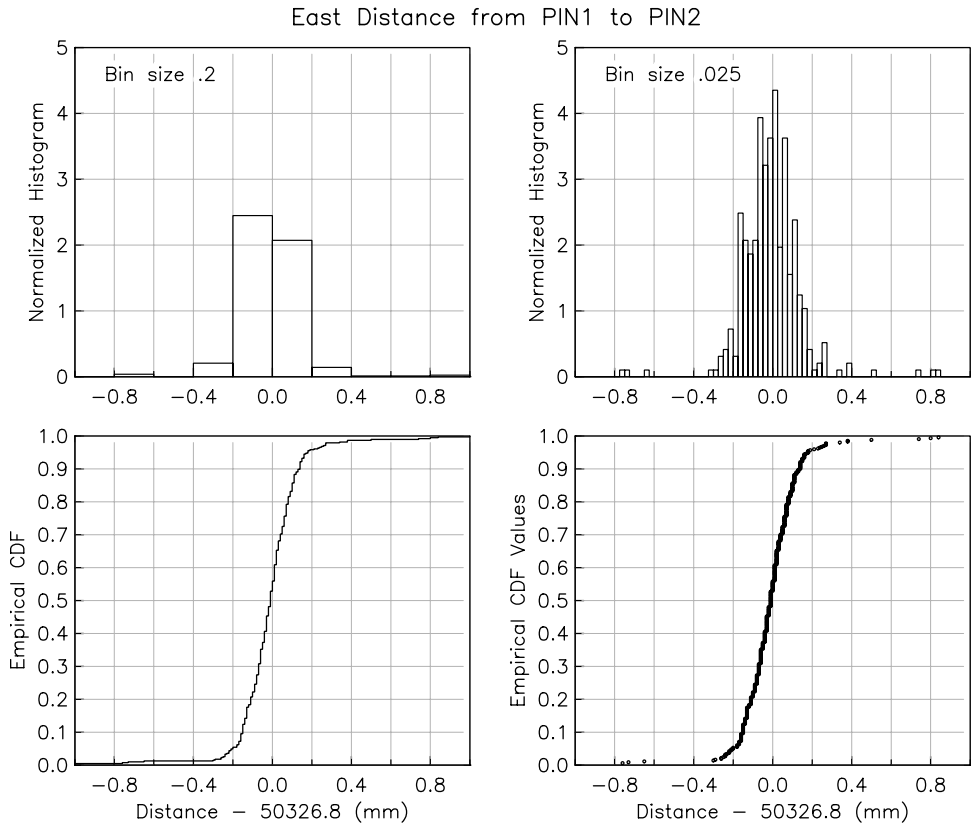

Figure 2.2: The top panels show histograms of the GPS data also plotted in Figure 1.1, for varying bin widths. The histograms are normalized so thapt they integrate to one. The lower left panel shows the empirical cumulative distribution function; the lower right panel shows the same information, but plotted as individual data points.

appearance depends what bin size we choose. Make the bin too large, and we lose resolution; too small, and there are a lot of (possibly meaningless) fluctuations. Figure 2.2 illustrates the problem. The bin width of Figure 1.1 is between these, but there is no way to say that it is "right".

But we can avoid having to choose a bin size if we instead construct and plot $S_n(x)$. In keeping with the Lebesgue Decomposition Theorem, this is usually defined as being like the cdf of a discrete pdf: a "stairstep" function that is continuous except for steps at each data value, and increases monotonically from zero to one. Suppose we have n data; we sort these into increasing order to form the set $(x_{(1)}, x_{(2)}, \ldots, x_{(n)})$, where the subscripted par-

entheticals are the standard notion for sorted (ordered) data. From these ordered data we create $S_n(x)$:

$$S_n(x) = \begin{array}{cl} 0 & x < x_{(1)} \\ i/n & x_{(i)} \le x < x_{(i+1)}, i = 1, \ldots, n-1 \\ 1 & x_{(n)} \le x \end{array}$$

Figure 2.2, also shows $S_n(x)$ for the GPS data; it appears fairly smooth even with no binning.

2.6.3 Probability Plots and Q-Q Plots

The lower right panel of Figure 2.2 shows the data for the empirical cdf plotted, not as a function, but as the values of the individual data points. We can use a modification of this plot both to show how well (or poorly) the data distribution agrees with a model, and to compare the distributions of two sets of data. Such plots provide qualitative information, as part of what is called **exploratory data analysis**: these are graphical methods that are the first thing you should apply to a new dataset.

The first method is the **probability plot**, which we can create given both the empirical cdf, $S_n(x)$, of a dataset and the cdf, $\Phi(x)$ of a particular probability distribution. Imagine warping the x-axis so that the function $\Phi(x)$ becomes a straight line, and then plotting $S_n(x)$ – or actually, just plotting the positions of the corners at the values of the data points. That is, we find a set of quantiles based on the number of data, n:

$$a_i = \Phi^{-1}\left(\frac{i - \frac{1}{2}}{n}\right) i = 1, \ldots, n$$

so that the a_i divide the area under $\Phi(x)$ into $n+1$ areas, each exactly $1/(n+1)$. To make a probability plot, we plot the sorted data against these values, as n pairs $a_i, x_{(i)}$. If the pdf is a good description of the data these points would fall on a straight line.

The top two panels of Figure 2.3 show probability plots for our GPS data set, using the Normal distribution as the assumed distribution. The left-hand panel shows almost the full dataset; we have not shown all the data because then all that can be seen is that that there are a few very large outliers. Even with these omitted, the distribution bends away from the diagonal, showing that there are more values far from the mean than would be present in a Normal distribution. This property is expressed by saying

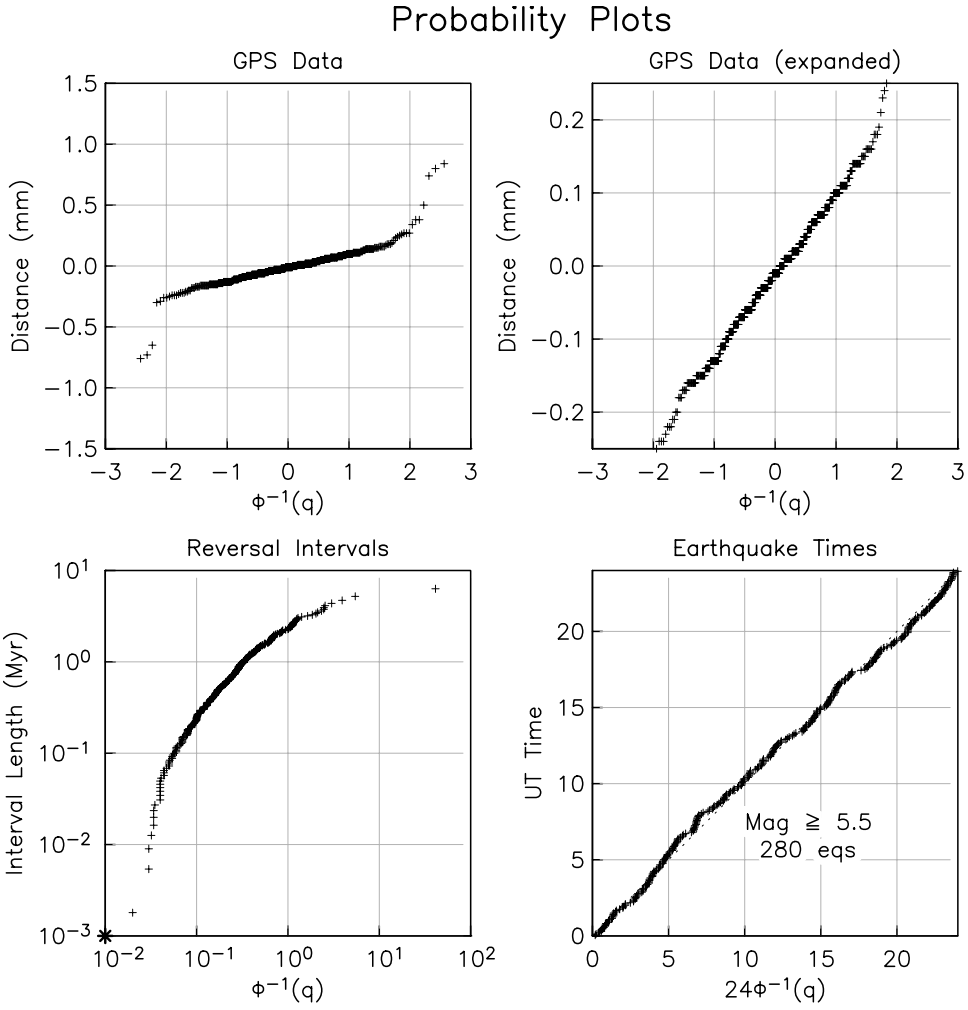

Figure 2.3: Probability plots for data sets from Chapter 1: for the GPS data compared with a Normal pdf, for the reversal data compared with an exponential distribution, and for the earthquake-time data compared with a uniform distribution.

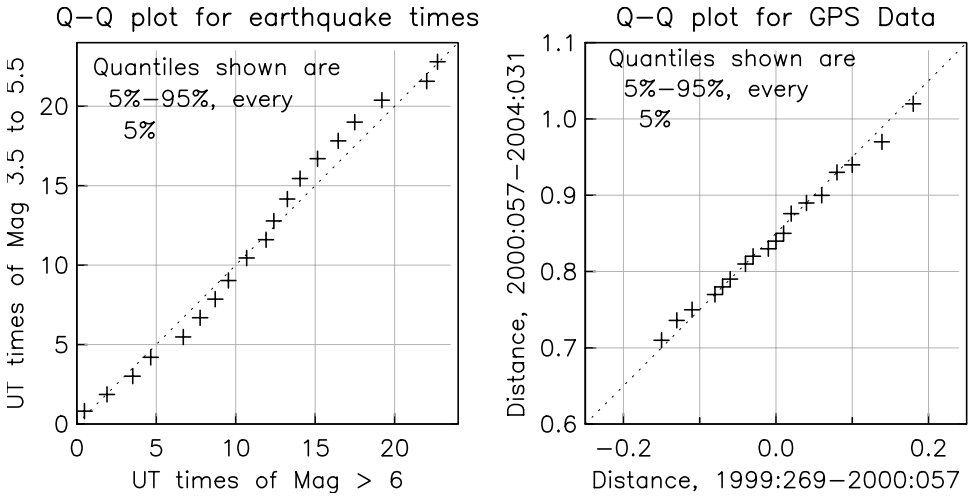

Figure 2.4:

that the actual distribution is **long-tailed** or **heavy-tailed**, which is not uncommon. The right-hand panel shows the distribution if we omit (that is, ignore) points beyond ± 0.25; the distribution then is impressively close to Normal. And the plot allows us to estimate several summary quantities that we will define in Section 2.7: the mean (from the zero intercept); the standard deviation (from the intercept of a line through the data, evaluated for $\Phi^{-1}(x) = \pm 1$), and the median and interquartile range.

The lower two panels show the other two datasets we use as examples. On the lower left we have the intervals between reversals, plotted on the assumption that the intervals are exponentially distributed (see Section 3.4.1) as they would be for a Poisson process. We have to use log scales on both axes to make the plot readable. The data clearly do not follow a straight line, so we may be sure that this model is not adequate. The lower right panel shows the times of earthquakes from Figure 1.5, plotted assuming that they are uniformly distributed. The event times seem to approximate uniformity fairly well, something we discuss more rigorously in Sections 6.3.1 and 6.5.1.

We can extend the plotting of quantiles to two datasets, making what is called a quantile-quantile or **Q-Q plot**. Suppose we have two sets of ordered data, of size n and m:

$$x_{(1)}, x_{(2)}, \ldots x_{(n-1)}, x_{(n)} \qquad \text{and} \qquad y_{(1)}, y_{(2)}, \ldots y_{(m-1)}, y_{(m)}$$

If $n = m$, then we simply plot the ordered pairs $(x_{(i)}, y_{(i)})$. If not, we have

to interpolate to get the quantiles, which we do by taking the quantile q, where $0 \leq q \leq 1$, and from it finding the values $r = qn + ½$ and $s = qm + ½$. We truncate these to get the integers k and l, and find the fractional parts $e = r - k$ and $f = s - l$. Then we can create interpolated "data values" that are (approximately) associated with this quantile:

$$x(q) = (1-e)x_{(k)} + ex_{(k+1)} \qquad \text{and} \qquad y(q) = (1-f)y_{(l)} + fx_{(l+1)}$$

We evaluate and plot these pairs $(x(q), y(q))$ for a selected set of q's; to make the plot useful it is usually best to use a finer sampling for q near 0 and 1, and a coarser sampling for q near ½. If we used this scheme with $n = m$, we could take q to be a multiple of n^{-1}, and would get, as we should, the ordered pairs $(x_{(i)}, y_{(i)})$. If the two data sets have the same distribution, the plotted points will lie along the line $y = x$. Shifts in the mean will move the points right or left from this line; differences in the variance will change the slope away from one. As with the probability plot, the advantage of the Q-Q plot is that it shows the behavior over the full range of the data, not merely a few summary values. And, it does not depend on any assumption about some "underlying" distribution: it comes from the data. However, we are usually limited to values of q between 0.05 and 0.95, so a Q-Q plot may not show if one of the datasets is longer-tailed than the other.

Figure 2.4 compares a couple of our "standard" datasets with closely related ones. The left panel shows the quantiles for times of large earthquakes against those of smaller ones;[5] the smaller ones have times that are nearly uniformly distributed, so the plot looks very much like the probability plot in Figure 2.3. On the right, we compare our GPS data with data for a later time span; there is a clear shift in location, and the later data (along the y-axis) having just slightly less variance than the earlier data, though a very similar pdf.

2.7 From RV's to Conventional Variables I: Summarizing PDF's

We now leave data behind (for the moment) and return to discussing random variables and the functions that describe them. Often it is useful to

[5] The smaller earthquakes are those between magnitude 3.5 and 5.4 in the Southern California earthquake catalog between 1981.0 and 2003.5, omitting days with 5.5 and larger shocks.

summarize certain attributes of a random variable, for example its "typical" value, or its variability; these attributes are conventional variables and not rv's. We can extract different kinds of summary variables from a pdf: we may lose information but computations are often more manageable when we consider summary variables rather than full functions.

The operations we perform on the pdf to get summary values are performed on the pdf, *not* on data – though operations on pdf's can suggest how we might summarize datasets, as we discuss in Chapter 5.

One possible summary of the typical value of a random variable, also called its "location", is the value of x which maximizes $\phi(x)$; this value is the **mode**. Though density functions often have one peak (unimodal), they may have many (multimodal, as in Figure 2.1), in which case the mode is not very useful. More generally, the mode can vary substantially even with small changes around the peak of the distribution: so it is a poor measure of the location of the rv.

A better summary value for the location comes from taking integrals or sums of the pdf to get the **mean** of the rv, which we symbolize by μ. If X is a continuous random variable with pdf $\phi(x)$ the mean comes from an integral whose integrand includes the pdf:

$$\mu = \int_{-\infty}^{\infty} x\phi(x)\,dx \tag{2.10}$$

where we can integrate over the entire real line because, if X is confined to only part of it, over the rest $\phi(x) = 0$. For a discrete random variable X, with probability distribution $p_i = \Pr[X = x_i], i = 1, 2, \ldots$; we can convert to a continuous rv using delta functions, with the pdf being $\sum_i p_i \delta(x - i)$). The mean of the rv is then

$$\mu = \sum_i p_i x_i$$

The other summary variable that is useful to have is one that expresses the spread of the rv, or equivalently the width of its pdf. One such variable is called the **variance** σ^2 and is computed from the pdf by

$$\mathcal{V}[X] \stackrel{\text{def}}{=} \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 \phi(x)\,dx \quad \text{or} \quad \mathcal{V}[X] \stackrel{\text{def}}{=} \sigma^2 = \sum_i p_i (x_i - \mu)^2 \tag{2.11}$$

where the first expression is how we write the variance of a random variable X; $\stackrel{\text{def}}{=}$ is shorthand for "the left side is defined to be what is on the right side".[6] $\mathcal{V}$ can be viewed as an operator that acts on a random variable

[6] Though sometimes we will write the defined quantity on the other side.

and produces a conventional one. For a Normal distribution,[7] $\sigma = \sqrt{\mathcal{V}[X]}$ is called the **standard deviation** or **standard error**, but this term is best avoided unless that particular distribution is being used or assumed. As we will see in Chapter 3, some pdf's are completely specified if we know μ and σ^2, but others are not.

Other measures of central value and spread are based on the quantiles of the cdf; as we showed above, the q-th quantile of the distribution Φ is the value x_q such that $\Phi(x_q) = \Pr[X \le x_q] = q$, so that $x_q = \Phi^{-1}(q)$. The **median** of the distribution, μ_m, is the quantile corresponding to $q = \frac{1}{2}$; this means that

$$\int_{-\infty}^{\mu_m} \phi(x)\,dx = \int_{\mu_m}^{\infty} \phi(x)\,dx = \frac{1}{2}$$

which is equivalent to $\Pr[X < \mu_m] = \Pr[X \ge \mu_m] = \frac{1}{2}$. The quantiles corresponding to $q = 0.25$ and $q = 0.75$ are called the lower and upper **quartiles** of Φ, and give another measure of spread; the difference $x_{0.75} - x_{0.25}$, which is known as the **interquartile range** or **IQR**. The dashed lines in Figure 2.1 show how the median and interquartile range are found for a particular cdf; in this case, with a multimodal pdf, neither the mean nor the mode are good values of location – actually, with a pdf shaped like this we cannot really summarize it very well with only one or two numbers.

Even small changes in the tails of a pdf can strongly influence the mean and variance, but these will have little effect on the median and interquartile range; we will see an example of this in Section 5.2. A summary value that is insensitive to small changes in the pdf is called **robust**. A robust measure of spread not based on quantiles is the the **mean deviation**, σ_m:

$$\sigma_m = \int_{-\infty}^{\infty} |x - \mu| \phi(x)\,dx$$

2.8 From RV's to Conventional Variables II: Moments

The mean and variance are examples of the **moments** of a pdf; these involve multiplying the pdf $\phi(x)$ by a power of x, and integrating or summing[8]. The mean is also called the **first moment**, and the variance is

[7] Note that, as is common, we say "probability distribution" even when we refer to the density function.

[8] The term "moment" comes from mechanics – remember moment of inertia.

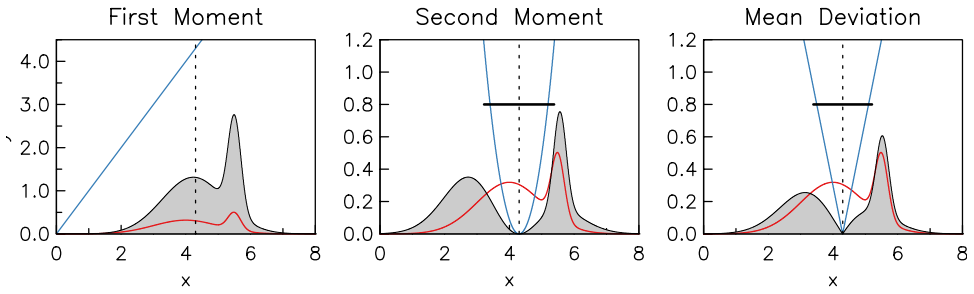

Figure 2.5: Graphical demonstration of the computation of the first moment (left panel, value is $\mu = 4.3$, which is the dashed line); the second centered moment (center panel, value is $\sigma^2 = 1.17$, $\sigma = 1.08$; the heavy line is $\mu \pm \sigma$); and the mean deviation (right panel, value is $\sigma_m = 0.91$; the heavy line is $\mu \pm \sigma_m$). In each panel the red line is the pdf, the blue line is the influence function, and the shaded region is the product, whose area gives the moment or deviation.

related to the **second moment**. There are actually two kinds of moments. The first kind, the moments about the origin, are

$$\mu'_r = \int_{-\infty}^{\infty} x^r \phi(x)\, dx \qquad \text{or} \qquad \mu'_r = \sum_i x_i^r p_i \tag{2.12}$$

where the expressions on the left and right give the r-th moments about the origin for $r = 1, 2, \ldots$ (why is $r = 0$ uninteresting?). The second kind, the moments about the mean, which are what is usually meant by "moment" for $r \geq 2$, are defined by

$$\mu_r = \int_{-\infty}^{\infty} (x - \mu'_1)^r \phi(x)\, dx \qquad \text{or} \qquad \mu_r = \sum_i (x_i - \mu'_1)^r p_i$$

for $r = 2, \ldots$. (Why is μ_1 uninteresting?) So the mean μ is μ'_1; the variance σ^2 is $\mu_2 = \mu'_2 - \mu^2$.

The moments of order higher than two of a pdf $\phi(x)$ provide additional summarizing information about the rv X. The third moment, μ_3, is a measure of the asymmetry of $\phi(x)$, also called the **skewness** of X; μ_4, known as the **flatness** (or **kurtosis**) further describes the shape of the pdf.

Figure 2.5 shows a graphical way of viewing computation of the first two moments, and also of the mean deviation. All of these can be viewed as the result of forming the product of $\phi(x)$ with some other function, and integrating the result. For example, for the first moment the multiplication

results in a new function that is (very roughly) the pdf scaled by the value of x around the location of the pdf; since the integral of $\phi(x)$ is one, the integral of $\phi(x)$ scaled to $c\phi(x)$ is just c. Similarly, the definition of the variance is the integral of $\phi(x)$ multiplied by a parabola centered on $\phi(x)$, and the mean deviation is the integral of $\phi(x)$ multiplied by two lines centered on $\phi(x)$. Both the parabola and the straight lines are examples of **influence functions**, so called because they give varying weight, or influence, to different parts of $\phi(x)$. Clearly the parabola gives more influence to the pdf far from its center – which may not be a good thing to do.

2.9 From RV's to Conventional Variables III: Expectations

Moving towards even greater generality, we come to an even more inclusive way of getting a conventional variable from a random one, again by integrating over an expression that includes the pdf. Suppose we have a function (strictly speaking, a functional) which maps the domain of the random variable into some other domain (for example, maps the real line into itself); we call this function g. When g operates on a random variable X, the result $Y = g(X)$ is another random variable. The **expected value** of $Y = g(X)$, also called its **expectation**, is given by

$$\int_{L_b}^{L_t} g(x)\phi(x)\,dx \tag{2.13}$$

where the limits are those applicable to $g(X)$; for example, if X could take on any positive or negative value, and $g(x)$ was x^2, the limits for g, and the integration, would be from zero to infinity. We write the expectation of a random variable Y as $\mathscr{E}[Y]$.

$\mathscr{E}$ is a kind of operator, like the variance operator $\mathscr{V}$ introduced in equation (2.11) – or like differentiation or integration. The expectation takes any random variable and creates a conventional variable from it; for a conventional variable c, $\mathscr{E}[c] = c$.[9] Because $\mathscr{E}$ involves integration, it is linear, so that

$$\mathscr{E}\left[\sum_{i=1}^{k} c_i g_i(X)\right] = \sum_{i=1}^{k} c_i \mathscr{E}[g_i(X)] \tag{2.14}$$

[9] One way to view this is that c is in fact a random variable C whose pdf is $\delta(x-c)$: so C is always equal to c

The simplest case is when $g(Y) = X$; then

$$\mathscr{E}[X] = \int_{L_b}^{L_t} x\phi(x)\,dx = \mu$$

where the last part comes from the definition of the mean: that the mean is $\mathscr{E}[X]$ is of course the reason for the name "expected value". Similarly, equations 2.10, 2.11, and 2.12 become

$$\mathscr{E}[X] = \mu \qquad \sigma^2 = \mathscr{V}[X] = \mathscr{E}[(X-\mu)^2] \qquad \mathscr{E}[X^r] = \mu'_r \tag{2.15}$$

2.10 Transformations and Functions of Random Variables

We now look at what we might call the arithmetic of random variables. Suppose that we produce a new random variable Y from a random variable X with pdf $\phi_X(x)$, how do we specify rules that tell us the pdf $\phi_Y(x)$ of Y. In the next section we will see that it is complicated to find the pdf of even the sum of two rv's, so we start with two simpler cases: first, combining rv's with conventional variables, and second, functions of a random variable.

A general combination of a random variable with conventional variables is a linear transformation, involving both multiplication and addition:

$$Y = c(X + l) \tag{2.16}$$

where we label the variables c and l because we will use these in Chapter 3 for the spread and location parameters of a pdf. Now consider the probability

$$\Pr[y \le Y \le y + g] = \int_{y}^{y+g} \phi_Y(v)\,dv \tag{2.17}$$

From equation (2.16) we have that

$$\Pr[y \le Y \le y + g] = \Pr[y \le c(X + l) \le y + g] \tag{2.18}$$

remembering that y, being the one end of a range, is just a conventional variable, and does not change when the random variable does. We can rewrite the right-hand side of equation (2.18) as

$$\Pr\left[\frac{x-l}{c} \le X \le \frac{x-l+g}{c}\right] = \int_{\frac{x-l}{c}}^{\frac{x-l+g}{c}} \phi_X(u)du \tag{2.19}$$

by the definition of the pdf ϕ_X. In order to make the limits on the integral the same as those in equation (2.17), we have to perform a change of variables, with $w = cu + l$ so $u = (w - l)/c$. Making this change, the integral in equation (2.19) becomes

$$\int_y^{y+g} \phi_X\left(\frac{w-l}{c}\right)\frac{dw}{c}$$

which means that

$$\phi_Y(x) = \frac{1}{c}\phi_X\left(\frac{x-l}{c}\right)$$

This is a result we will use frequently, for example, in Section 3.3.1.

Intuitive though the pdf is, the result derived above can be found much more easily if we use the cumulative distribution function instead. We can put most of the steps on one line:

$$\Phi_Y(y) = \Pr[Y \le y] = \Pr[cX + l \le y] = \Pr\left[X \le \frac{y-l}{c}\right] = \Phi_X\left(\frac{y-l}{c}\right) \qquad (2.20)$$

and have only to take the derivatives:

$$\phi_Y(y) = \frac{d}{dy}\Phi_Y = \frac{d}{dy}\Phi_X\left(\frac{y-l}{c}\right) = \frac{1}{c}\phi_X\left(\frac{y-l}{c}\right)$$

where the c^{-1} in the last expression comes from the chain rule for derivatives.

Now suppose we have a more general case, in which $Y = g(X)$; we assume that over the range of X; $g(X)$ is differentiable, and also monotone, so that there is an inverse function that satisfies $X = g^{-1}(y)$. Then we can use the same approach to relate $\phi_Y(y)$ to $\phi_X(x)$. We follow the steps in equation (2.20) and write

$$\Phi_Y(y) = \Pr[Y \le y] = \Pr[g(X) \le y] = \Pr[X \le g^{-1}(y)] = \Phi_X(g^{-1}(y)) \qquad (2.21)$$

which we differentiate, using the chain rule, to get

$$\phi_Y(y) = \frac{d}{dy}\Phi_Y = \frac{d}{dy}\Phi_X(g^{-1}(y)) = \phi((g^{-1}(y))\left|\frac{d}{dy}(g^{-1}(y))\right| \qquad (2.22)$$

where the absolute value is present to deal with the case in which Y decreases as X increases.

As an example, suppose that we have $\phi = 1$ for $0 \le X \le 1$ (the uniform distribution) and want the pdf of $Y = X^2$. Then $g^{-1}(y) = \sqrt{y}$, and

$$\phi_Y(y) = \frac{1}{2\sqrt{y}}$$

which is interesting because it shows that the pdf can be infinite, provided only that the associated singularity is integrable.

While 2.22 might appear to provide a simple formula to apply, it is actually better in practice to start with the steps in 2.21, which are more general and easier to remember. If, for example, we had $\phi = 1$ for $-\frac{1}{2} \le X \le \frac{1}{2}$ and $Y = X^2$, we could not use 2.22 because there is no unique inverse; but the steps in 2.21 become

$$\Phi_Y(y) = \Pr[Y \le y] = \Pr[X^2 \le y] = \Pr[-\sqrt{y} \le X \le \sqrt{y}] = \Phi_x(\sqrt{y}) - \Phi_x(-\sqrt{y})$$

from which the pdf, $y^{-\frac{1}{2}}$ for $0 \le y \le 0.25$, can easily be derived.

2.11 Sums and Products of Random Variables

We have not, so far, dealt with such basic arithmetic operations as adding or multiplying two variables together. So the next question we address is, given two rv's X_1 and X_2 with known pdf's, what are the pdf's of $X_1 + X_2$, $X_1 X_2$, and X_1/X_2? We derive some results for these combinations in this section, we can then use the one for summation to demonstrate the Central Limit Theorem (Section 2.12). We will also use the results of this section extensively in Chapter 3 in the course of deriving a variety of pdf's.

2.11.1 Summing Two Variables

Our first step may appear to just complicate the problem; this is to generalize the concept of a pdf to more than one variable. This generalization requires us to introduce the concept of **joint probability** for random variables. We already have joint probabilities for sets: the joint probability for set A and set B is $\Pr[A \cap B]$. If we say that set A is having X_1 fall between x_1 and $x_1 + \delta x_1$, and set B is having X_2 fall between x_2 and $x_2 + \delta x_2$, then

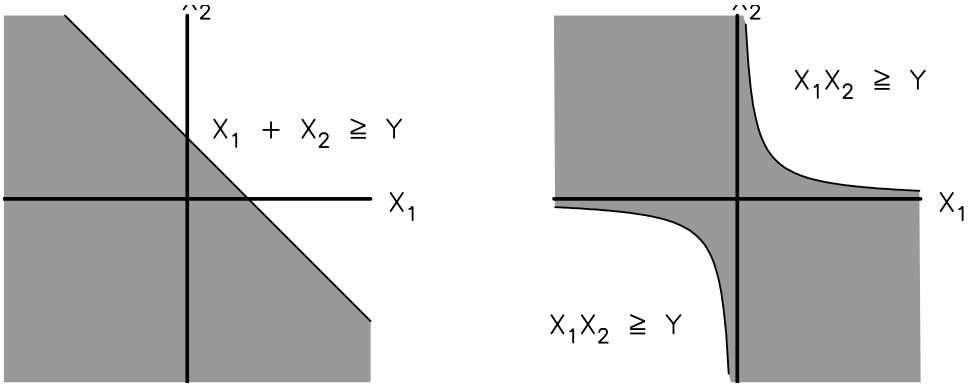

Figure 2.6: In the left panel the shaded area is the region for the integral in equation (2.24) for finding the pdf of the sum of two rv's. In the right panel the shaded area is the region for the integral in equation (2.25) for finding the pdf of the product of two rv's.

we can write the joint probability in terms of a pdf of two variables:

$$\Pr[(x_1 \le X_1 \le x_1+\delta x_1)\cap(x_2 \le X_2 \le x_2+\delta x_2)] = \int_{x_1}^{x_1+\delta x_1}\int_{x_2}^{x_2+\delta x_2} \phi(x_1,x_2)\,dx_1 dx_2 \tag{2.23}$$

which we write as $X_1, X_2 \sim \phi(x_1, x_2)$, meaning that the random variables X_1 and X_2 are **jointly distributed** with pdf $\phi(x_1, x_2)$.

To find the pdf for the sum, we introduce the rv $Y = X_1 + X_2$, which has the pdf ψ and distribution Ψ. Then we find

$$\Psi(y) = \Pr[Y \le y] = \Pr[X_1 + X_2 \le y] = \int_{x_1+x_2\le y} \phi(x_1,x_2)\,dx_1 dx_2 \tag{2.24}$$

so that the integral is over the shaded area on the left of Figure 2.6.

To proceed beyond this, we have to assume that the random variables X_1 and X_2 are independent; Chapter 4 will deal with the case that they are not. For sets independence means that $\Pr[A \cap B] = \Pr[A]\Pr[B]$; this can be consistent with equation 2.23 only if the pdf for the two variables has the form

$$\phi(x_1,x_2) = \phi_1(x_1)\phi_2(x_2)$$

where ϕ_i is the pdf of X_i. In this case the properties of X_1 can be found independently of the distribution of X_2; that is to say, from ϕ_i alone. Then

$$\Psi(y) = \int_{x_1+x_2\le y} \phi_1(x_1)\phi_2(x_2)\,dx_1 dx_2$$

Letting $s = x_1 + x_2$ we get

$$\Psi(y) = \int_{-\infty}^{y}\int_{-\infty}^{\infty} \phi_1(x_1)\phi_2(s-x_1)\,dx_1 ds$$

Differentiating gives the pdf for Y:

$$\psi(y) = \frac{d\Psi}{dy} = \int_{-\infty}^{\infty} \phi_1(x_1)\phi_2(y-x_1)\,dx_1 \stackrel{\text{def}}{=} \phi_1 * \phi_2$$

In the last part of this equation we have introduced a new notation, namely $*$ to mean the particular integral of a product of functions, which is called the **convolution** of the two functions ϕ_1 and ϕ_2 to form the function ψ. We can generalize this result for multiple independent rv's X_1, X_2, ..., X_n, with $X_k \sim \phi_k$: the sum has the pdf

$$X_1 + X_2 + \ldots + X_n \sim \phi_1 * \phi_2 * \phi_3 \ldots * \phi_n$$

which is to say, if we add independent random variables, we get a random variable whose pdf is the convolution of the component pdf's.

2.11.2 Multiplying Two Variables

For the product of two rv's, we proceed similarly to the derivation for sums: we introduce the rv $Y = X_1X_2$, with pdf ψ and distribution Ψ; $Y \sim \psi$ with $\psi = \frac{d\Psi}{dy}$. Then

$$\Psi(y) = \Pr[Y \le y] = \Pr[X_1X_2 \le y]$$

To get this, we have to integrate the joint pdf $\phi(x_1, x_2)$ over the set such that $x_1x_2 \le y$; if $x_1 \ll 0$, $x_2 \ge y/x_1$, while if $x_1 \gg 0$, $x_2 \le y/x_1$, making the integral of the joint pdf over the shaded area on the right of Figure 2.6. We can write this as the sum of two integrals

$$\int_{-\infty}^{0}\int_{y/x_1}^{\infty} \phi(x_1,x_2)\,dx_2dx_1 + \int_{0}^{\infty}\int_{-\infty}^{y/x_1} \phi(x_1,x_2)\,dx_2dx_1 \tag{2.25}$$

We introduce a new variable $s = x_1x_2$, which makes 2.25

$$\begin{aligned}
&\int_{-\infty}^{0}\int_{y}^{-\infty} \frac{1}{x_1}\phi(x_1,s/x_1)\,dsdx_1 + \int_{0}^{\infty}\int_{-\infty}^{y} \frac{1}{x_1}\phi(x_1,s/x_1)\,dsdx_1 \\
&= \int_{-\infty}^{0}\int_{-\infty}^{y} \frac{1}{-x_1}\phi(x_1,s/x_1)\,dsdx_1 + \int_{0}^{\infty}\int_{-\infty}^{y} \frac{1}{x_1}\phi(x_1,s/x_1)\,dsdx_1 \\
&= \int_{-\infty}^{y}\int_{-\infty}^{\infty} \frac{1}{|x_1|}\phi(x_1,s/x_1)\,dx_1 ds
\end{aligned}$$

Since this is $\Psi(Y)$, we can differentiate to get

$$\psi(y) = \int_{-\infty}^{\infty} \frac{1}{|x_1|} \phi(x_1, y/x_1)\, dx_1 = \int_{-\infty}^{\infty} \frac{1}{|x_1|} \phi_1(x_1)\phi_2\left(\frac{y}{x_1}\right) dx_1 \tag{2.26}$$

where only at the last step have we assumed that X_1 and X_2 are independent. A similar approach for $Y = X_1/X_2$ gives

$$\psi(y) = \int_{-\infty}^{\infty} |x_1| \phi_1(x_1)\phi_2(x_1 y)\, dx_1 \tag{2.27}$$

which we will also use in Chapter 3.

2.12 The Central Limit Theorem

In probability theory and statistical inference the **Normal distribution**, also called the **Gaussian distribution**, plays a major role. The pdf for this, with the mean set to zero, is

$$\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2}$$

which is conventionally written as $X \sim N(0, \sigma)$.

This major role importance of the normal distribution is in part justified by the **central limit theorem**; loosely speaking this theorem states that if a random variable X is the sum of a large number of other random variables, then X will be approximately normally distributed, irrespective of the pdfs of the variables that are summed to produce it.

We now demonstrate this, with two caveats. The first is that this is not a fully rigorous proof; pp. 113-116 of **?** has one. The second is that, despite this theorem, actual data (see, for example, Figure 1.1) have a stubborn habit of *not* being normally distributed. Often they are "close enough" that it doesn't matter (much), but your should *always* check this, and allow for the possibility that the data are not normal.

2.12.1 The Characteristic Function

We begin our demonstration by doing something that, unless you are already familiar with convolution, will not be obvious: we take the Fourier transform of the pdf. We do so because the convolution operation on the pdf's is then replaced by multiplication of their Fourier transforms, which

is much more manageable. For a random variable $X \sim \phi$, the Fourier transform of ϕ is written as $\mathscr{F}[\phi]$ or $\tilde{\phi}(f)$, and defined as

$$\tilde{\phi}(f) = \int_{-\infty}^{\infty} \phi(x) e^{-2\pi i f x}\, dx$$

In probability theory this transform is called the **characteristic function** of ϕ; it has the inverse transform

$$\phi(x) = \int_{-\infty}^{\infty} \tilde{\phi}(f) e^{2\pi i f x}\, df$$

Because pdf's are such well-behaved functions (positive and with a finite area) these two transforms always exist. Note that

$$\tilde{\phi}(0) = \int_{-\infty}^{\infty} \phi(x)\, dx = 1$$

where the integral follows from direct substitution into the Fourier-transform equation.

The inverse Fourier transform shows that the characteristic function uniquely determines the pdf, and so, just as much ϕ itself does, completely specifies the properties of X; so we can use this function for demonstrating theorems. To start, note that the characteristic function can also be defined in terms of the expectation operator (equation 2.13); if we take $g(x)$ to be $e^{-2\pi i f x}$, we see that the Fourier transform corresponds to our definition of an expectation, so that

$$\tilde{\phi}(f) = \mathscr{E}\left[e^{-2\pi i f X}\right] \tag{2.28}$$

which, mysterious as it might first appear, is just the application of a function to some random variable.

Expanding the exponential in equation (2.28), and making use of the linearity of the expectation operator (equation 2.14) and the definition of the higher moments (equation 2.12), gives a Taylor series expansion of the characteristic function:

$$\tilde{\phi}(f) = \sum_{r=0}^{\infty} \frac{(-2\pi i f)^r}{r!} \mu'_r \tag{2.29}$$

where $\mu'_r = \mathscr{E}[X^r]$. Since a Taylor series determines the function (and hence the pdf) uniquely, we have shown that, given all its moments μ', a pdf is completely determined.

We can use equation (2.29) to express the mean and variance of a distribution in terms of derivatives of the characteristic function, evaluated at zero. For example, if we take the derivative of equation (2.29), and then evaluate it at zero, we have only one term left in the expansion, so that

$$\tilde{\phi}\prime(0) = -2\pi i \mu_1' \qquad \text{whence} \qquad \mathscr{E}[X] = \frac{-\tilde{\phi}'(0)}{2\pi i}$$

by equation (2.15); similarly, $\tilde{\phi}''(0) = -4\pi^2\mu_2'$, so for the variance we get

$$\mathscr{V}[X] = \frac{\tilde{\phi}'(0)^2}{4\pi^2} - \frac{\tilde{\phi}''(0)}{4\pi^2}$$

so, since $\tilde{\phi}''(0) \le 0$, $\mathscr{V}[X] \ge 0$, as it should be.

2.12.2 Summing Many Variables

For demonstrating the Central Limit Theorem we first find the characteristic function of the Normal pdf:

$$\tilde{\phi}(f) = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2/\sigma^2)} e^{-2\pi i f x}\, dx$$

This integral may be evaluated by completing the square in the exponent to give a definite integral in x, which then gives another Gaussian:

$$\tilde{\phi}(f) = e^{-2\pi^2\sigma^2 f^2}$$

From this we can already see that if we sum n Gaussian random variables the resulting characteristic function $\tilde{\phi}_n$ will be

$$\tilde{\phi}_n(f) = e^{-2\pi^2 n\sigma^2 f^2}$$

and undoing the FT yields

$$\phi_n(x) = \frac{1}{\sqrt{2\pi n}\sigma} e^{-\frac{1}{2}(x^2/n\sigma^2)}$$

Now, we relax the assumption of a Gaussian distribution and just suppose that we have random variables $X_1, X_2, \ldots X_n$, which are **independent and identically distributed**, an assumption so common that it gets its

own acronym, namely **iid**. We assume the pdf has a mean of zero, a variance σ^2, and that all the higher moments exist.[10] Let $S_n = \sum_{i=1}^{n} X_i$. The Central Limit Theorem is the statement that, in the limit as $n \to \infty$, the distribution of S_n approaches $N(0, \sigma\sqrt{n})$; the variance σ^2 grows as n. If $S_n \sim \phi_n$ and each $X_i \sim \phi$ then ϕ_n is an n-fold convolution

$$\phi_n = \phi * \phi * \phi * \ldots * \phi$$

which means that the characteristic function $\tilde{\phi}_n$ is given by

$$\tilde{\phi}_n = \tilde{\phi} \cdot \tilde{\phi} \cdot \ldots \tilde{\phi} = (\tilde{\phi})^n = e^{n \ln \tilde{\phi}}$$

Assuming that all the moments of ϕ exist, then so do all the derivatives of $\tilde{\phi}$ at $f = 0$ and we can expand $\tilde{\phi}$ in a Taylor series:

$$\begin{aligned}\tilde{\phi}(f) =& \tilde{\phi}(0) + \frac{f}{1!}\tilde{\phi}'(0) + \frac{f^2}{2!}\tilde{\phi}''(0) + \ldots \\ =& 1 + \frac{f^2}{2!}\tilde{\phi}''(0) + \frac{f^3}{3!}\tilde{\phi}'''(0) + \ldots\end{aligned}$$

where we have made use of $\tilde{\phi}(0) = 1$ (true for all $\tilde{\phi}$) and $\tilde{\phi}'(0) = 0$ (because we assumed $\mathcal{E}[X] = 0$).

Putting this series into the $e^{n \ln \tilde{\phi}}$, we get

$$\begin{aligned}\tilde{\phi}_n(f) &= \exp\left[n \ln\left(1 + \frac{f^2}{2!}\tilde{\phi}''(0) + \frac{f^3}{3!}\tilde{\phi}'''(0) + \ldots\right)\right] \\ &\approx \exp\left[\frac{nf^2}{2!}\tilde{\phi}''(0) + \frac{nf^3}{3!}\tilde{\phi}'''(0) + \ldots\right]\end{aligned}$$

The approximation in the second line above has used the Taylor-series expansion for $\ln(1+\epsilon) = \epsilon - \epsilon^2/2 + \epsilon^3/3 \ldots$, keeping only the linear term in ϵ. Next we define a new variable

$$\sigma_n^2 = \mathcal{V}[S_n] = \frac{-n\tilde{\phi}''(0)}{4\pi^2} = n\mathcal{V}[X_i]$$

and a constant

$$c_3 = \frac{4\tilde{\phi}'''(0)}{3(-\tilde{\phi}''(0)/\pi)^{3/2}}$$

[10] In Section 3.5.1 we will encounter a pdf for which these moments do not in fact exist.

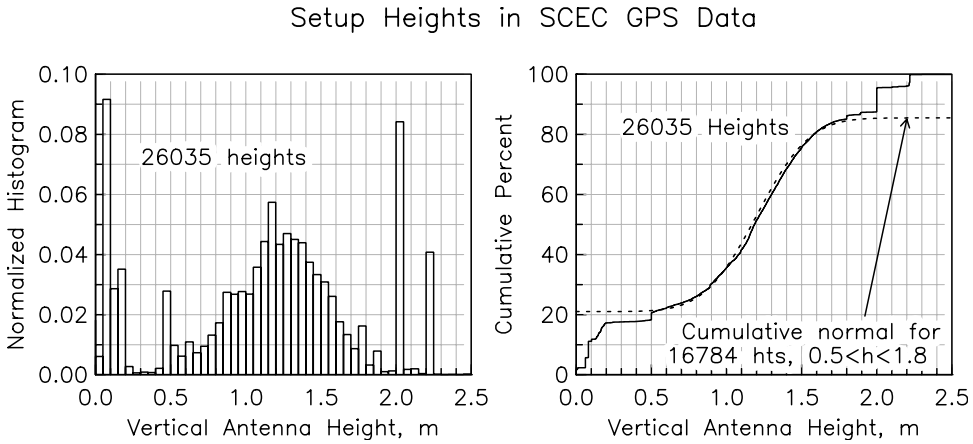

Figure 2.7: The histogram (left) and the cumulative distribution function (right) for the heights at which GPS antennas were set above the ground, for a very large database. Heights outside the range from 0.5 to 1.8 m usually involved some kind of pole with a fixed height; between these heights the usual stand for the antenna was a surveyor's tripod, which can be set up over a wide range. The dashed line in the right-hand plot is the cdf for a Normal distribution with mean 1.204 m and standard deviation 0.275 m.

in terms of which we can rewrite the series as

$$\exp\left[-2\pi^2\sigma_n^2 f^2 + \frac{(\sigma_n f)^3 c_3}{\sqrt{n}} + O\left(\frac{(\sigma_n f)^4}{n}\right)\right]$$

The effect of introducing σ has been to make all terms but the first approach zero as $n \to \infty$, and the first term gives a Gaussian characteristic function, with $\mathcal{V}[S_n] = n\mathcal{V}[X_i]$; $\phi(x)$ tends to a Gaussian with mean zero and variance $n\mathcal{V}[X] = n\sigma^2$, which is what we wanted to show.

To go from mathematics to real data, Figure 2.7 shows the histogram, and cumulative distribution function for a collection of very independent data: setups of surveyors' tripods, representing nearly 17,000 decisions by hundreds of people over two decades. In the right frame, the dashed line shows that, indeed, over a wide range the empirical cumulative distribution function for these heights is very nearly a Gaussian, or Normal, distribution, with only two parameters needed to describe a large dataset.