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 (1977) *The Advanced Theory of Statistics*, Chapter 7.

2. Introduction

This chapter introduces 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, namely the random variable, including the probability density function and distribution function that describe such a variable. We then define means, variances, expectations, and moments of these functions, and discuss, more briefly, distributions of more than one variable, which leads to the idea of independence. We close with the central limit theorem, which is a major justification for using the Normal (or Gaussian) distribution.

2.1. What is Probability?

As we noted in Chapter 1, there is dispute over what things in the real world the formal mathematical system of probability theory corresponds to. The two usual views can be briefly stated as

- The **frequentist** interpretation: the probability of something corresponds to what fraction of the time it happens "at random" or "in the long run". This is might also be called the casino interpretation of probability, since that is one setting where it seems to make sense; but there are many others in which it does not. Geophysics has many of these: 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 the Earth's gravitational field: there is only one example of this.
- The **Bayesian** or **subjective** interpretation, in which the probability of something corresponds to how likely we think it is to happen. In this approach, probabilities represent states of mind. As we will see, this approach leads to a distinctive set of methods for analyzing data.

The approach we prefer is the one we have already hinted at: probability is a mathematical system, which can be used as a model of certain aspects of the real world, just as we use other mathematical idealizations: for example, in studying seismic waves, we represent the Earth by an elastic solid—equally a mathematical idealization. If we simply take probability as a model, it can represent more than one kind of thing, so both interpretations can be valid.

[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—which we discuss in section 2.1.

2.2. Basic Axioms

The basic idea of the mathematical theory of probability, as developed by Kolmogorov on the basis of set theory, is the idea of a **sample space** Ω, which is a set that contains as elements 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 numerical values.

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

1. $\mathcal{P}(\Omega) = 1$; the probability of all the outcomes combined is 1 (has to happen).
2. $\mathcal{P}(A) \geq 0$; probabilities are positive.
3. If two sets of outcomes are disjoint (mutually exclusive) then $\mathcal{P}(A_i \cup A_j) = \mathcal{P}(A_i) + \mathcal{P}(A_j)$: the probability of the combination (union of the sets) is the sum of the individual probabilities.[2] That is, if having A precludes B 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, indeed, almost intuitive. But we can, from these few axioms, produce a full theory.

2.3. Conditional Probability

Things become slightly more interesting (because less obvious) once we introduce the concept of **conditional probability**, which is written as $\mathcal{P}(A \mid B)$, 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 $\mathcal{P}(A \mid B)\mathcal{P}(B) = \mathcal{P}(A \cap B)$: 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:

$$\mathcal{P}(A \mid B) = \frac{\mathcal{P}(A \cap B)}{\mathcal{P}(B)} \tag{1}$$

From this we get the concept of two sets of events being **independent**: A and B are independent if $\mathcal{P}(A \mid B) = \mathcal{P}(A)$, which is to say that the probability of A does not depend on whether B has happened or not[3]. This means, from (1), that $\mathcal{P}(A \cap B) = \mathcal{P}(A)\mathcal{P}(B)$: the probability of having both A and B is the product of their individual probabilities. In practice, this rule is easily abused, since it is all too tempting to decide that events are independent when they actually are not.

2.3.1. An Application: Was That a Foreshock?

An application of conditional probabilities to a geophysical problem (indeed, one of actual social significance) is the question of what we should do if a small earthquake occurs close to (say) the San Andreas fault, given that it might either be a foreshock to a major earthquake on this fault, or it might just be a background shock that happened there by chance. The full treatment[4] becomes

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

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

[4] D. C. Agnew and L. M. Jones, Prediction probabilities from foreshocks, *J. Geophys. Res.*, **96**,

rather complicated, but a simplified version runs as follows. We define three possible events:

B: A background earthquake has occurred.

F: A foreshock has occurred.

C: A large (so-called characteristic) earthquake will occur.

Of course, if a small background shock were to happen by coincidence just before the characteristic earthquake, we would certainly class it as a foreshock. Thus, B and C cannot occur together: they are disjoint. 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 (1),

$$\mathcal{P}(C \mid F \cup B) = \frac{\mathcal{P}(C \cap (F \cup B))}{\mathcal{P}(F \cup B)} \tag{2}$$

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 as

$$\mathcal{P}((C \cap F) \cup (C \cap B)) = \mathcal{P}(C \cap F) + \mathcal{P}(C \cap B) = \mathcal{P}(C \cap F)$$

where the disjointness of C and B eliminates the $\mathcal{P}(C \cap B)$ term. Again using the definition of conditional probability,

$$\mathcal{P}(C \cap F) = \mathcal{P}(F \mid C)\mathcal{P}(C) \tag{3}$$

where $\mathcal{P}(F \mid C)$ is the probability that a mainshock is preceded by a foreshock. Again using the disjointness of F and B, we can write the denominator as

$$\mathcal{P}(F \cup B) = \mathcal{P}(F) + \mathcal{P}(B) \tag{4}$$

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

$$\mathcal{P}(F) = \mathcal{P}(F \cap C) = \mathcal{P}(F \mid C)\mathcal{P}(C) \tag{5}$$

We can use (3), (4), and (5) to write (2) as

$$\mathcal{P}(C \mid F \cup B) = \frac{\mathcal{P}(F)}{\mathcal{P}(F) + \mathcal{P}(B)} = \frac{\mathcal{P}(C)\mathcal{P}(F \mid C)}{\mathcal{P}(F \mid C)\mathcal{P}(C) + \mathcal{P}(B)} \tag{6}$$

For $\mathcal{P}(B) \gg \mathcal{P}(F \mid C)\mathcal{P}(C)$ this expression is small (the candidate event is probably a background earthquake), while for $\mathcal{P}(B) = 0$, the expression becomes equal to one: any candidate earthquake must be a foreshock.

The second form of expression in (6) is a function of three quantities, which in practice we obtain from very different sources. $\mathcal{P}(B)$, the probability of a background earthquake, would be found from seismicity catalogs for the fault zone. $\mathcal{P}(C)$, the probability of a characteristic earthquake, would be found from the past history of large earthquakes on this fault as found from paleoseismological studies. If we had a record of the seismicity before many such characteristic earthquakes, we could evaluate $\mathcal{P}(F \mid C)$ But, given the limited time over which there is seismicity data, we do not have such a record; we in practice assume that the average of $\mathcal{P}(F \mid C)$ over many earthquakes on one fault is equal to the spatial average over many faults over a shorter time; even though this may not be valid, it is the best we can do.

2.3.2. Natural Frequencies: Another Frame for the Problem

While the algebraic manipulations in Section 2.2.1 are needed to solve the full problem, this is not the easiest way to get the result at the level given there. It turns out that insight into problems of this sort depends very much on how they are phrased.[5] For almost everyone, stating the numbers in terms of probabilities does not help intuitive reasoning;

11959-11971 (1991).

[5] Hoffrage, Ulrich, Samuel Lindsey, Ralph Hertwig, and Gerd Gigerenzer (2000). Communicating statistical information, *Science* **290**, 2261-2262.

what works much better is to state them in terms of numbers of events (out of some large but arbitrary number), an approach called **natural frequencies**. We recommend this approach either for explaining conditional probability reasoning to other people, or to yourself: it is a good way to check your algebra. To show how it works for the above example: 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 an 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 (6) to get the same result.

2.4. Bayes' Theorem

The procedures followed for the foreshock probability estimate are basically those used to derive **Bayes' Theorem**, a basis for much statistical inference—though, as we have mentioned above and will discuss further, how much it should be, is a subject of considerable debate. The theorem itself is not difficult to derive. Suppose we have N *disjoint* sets of outcomes, called $B_1, \ldots, B_N$, and another set A. The the probability of both A and a particular one of the B's (say B_j) is, by the definition of conditional probability,

$$\mathcal{P}(A \cap B_j) = \mathcal{P}(B_j | A)\mathcal{P}(A) = \mathcal{P}(A \mid B_j)\mathcal{P}(B_j) \tag{7}$$

where you should remember that $\mathcal{P}(A \cap B_j) = \mathcal{P}(B_j \cap A)$. But, since the B's are disjoint, $\mathcal{P}(A) = \sum_j \mathcal{P}(A \mid B_j)\mathcal{P}(B_j)$. Combining this with (7), we find that

$$\mathcal{P}(B_j | A) = \frac{\mathcal{P}(A \mid B_j)\mathcal{P}(B_j)}{\sum_j \mathcal{P}(A \mid B_j)\mathcal{P}(B_j)} \tag{8}$$

The different parts of this expression have special names: $\mathcal{P}(B_j)$ is called the **prior probability** of B_j, and $\mathcal{P}(A \mid B_j)$ the **likelihood** of A given B_j.

All this is unproblematic; the contentiousness comes (as usual) in how this can be applied to reasoning about things in the real world. One application is to suppose that the B's are degrees of belief about something: for example, $\mathcal{P}(B_1)$ would be our belief (expressed as a probability) that a coin is fair, $\mathcal{P}(B_2)$ 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 (8) allows us to find $\mathcal{P}(B_j | A)$, the **posterior probability** of each hypothesis.

The attractiveness of this scheme is clear: we have used the data directly to improve our degree of belief in one or another fact about the world, which is what we would like to do with all data; this is called **Bayesian inference**. However, there is a problem which we have evaded: how to determine the prior probabilities. We have been evasive for good reason, namely that deciding on prior probabilities is a complicated and controversial matter. So for now we put Bayes' theorem and Bayesian inference aside.

2.5. Random Variables: Density and Distribution Functions

Up to now we have talked about "outcomes" which are described by set theory. But most of the time, the things we want to model are described by numbers, which leads us to

the idea of a **random variable**, which we call (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 call conventional variables. The random variable is a different 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 could 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. We will also (sometimes) use the abbreviation (common in the statistics literature) **rv** for random variable.

A common source of confusion is that these two kinds of variables can refer to very similar things in the world—though not identical. Consider (again) the paradigmatic case of rolling dice. For a particular roll of the dice, the conventional variable x describes what we actually got—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 (often abbreviating to **rv**) 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. But the mapping can be into different spaces of numbers. If Ω maps into some part of the real line $\mathcal{R}$ (or the whole of it) X is termed a **continuous** random variable; if it maps into the integers, the random variable is a **discrete** one (as in our dicing example). Either way, each element ω in Ω corresponds to a unique number $X(\omega)$.

To describe X we need a probabilistic description, which turns out to be a function, called the **probability density function** of X. We approach the idea of a density function by looking first at a very common way in which we express the relative frequency of observing different values of a random variable: the **histogram**. We have already seen examples of this in Chapter 1, in the distance between two GPS stations and the lengths of magnetic dipole states. In figures 1.1 and 1.2 we plotted the number of observations; we could make the plots more independent of the particulars of these datasets if we instead plotted, in each bin, the number in the bin divided by the total number of observations.

Such a normalized histogram is a crude estimate of a probability density function for the random variable in question. This function is often referred to by its acronym (**pdf**), and we will designate it as $\phi(x)$. Note how much more "complicated" this makes random variables compared to conventional ones: a conventional variable is completely specified by a number, while to specify a rv takes a function (which may have many parameters).

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 denote the probability of X lying in this interval as:

$$\mathcal{P}(x \le X \le x+\delta x) = \text{Prob}(x \le X \le x+\delta x) = p[x \le X \le x+\delta x].$$

where we have used two other common notations for probability, Prob(), and $p[]$. Then in equation form the property of the pdf is that

$$\mathcal{P}(x \le X \le x + \delta x) = \int_{x}^{x+\delta x} \phi(u)\, du$$

For any x and small interval δx this means

$$\mathcal{P}(x \le X \le x + \delta x) \approx \phi(x)\delta x + (\delta x)^2$$

so that $\phi(x)$ represents the density of probability per unit value of x in the neighborhood of x. Probability density functions must satisfy:

1. $\phi(x) \ge 0$ for all x: probabilities are always positive.
2. $\int_{L_b}^{L_t} \phi(x)\, dx = 1$: X must take on some value within its permissible range. Often this is the real line, with $L_b = -\infty$ and $L_t = \infty$; but sometimes it is only part of that line. For example, time intervals have to be positive, in which case $L_b = 0$ and $L_t = \infty$; if we are 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, $\mathcal{P}(X = x) = 0$ for all x: the probability of X being exactly some value is zero.

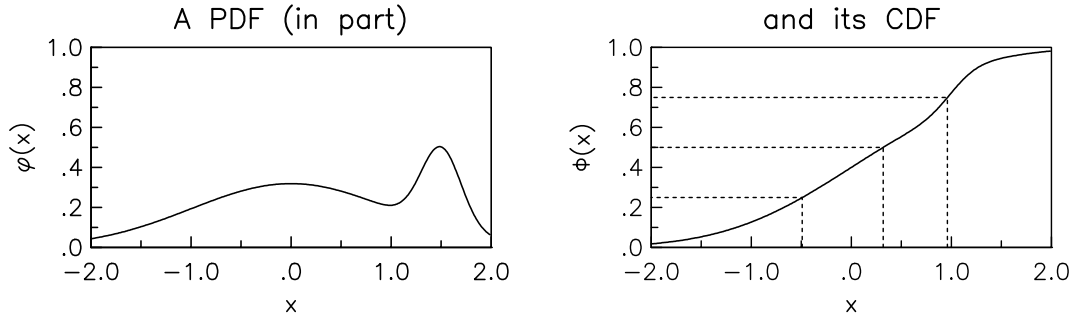

Figure 2.1

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

$$\Phi(x) = \int_{L_B}^{x} \phi(x) dx$$

which means that

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

provided this derivative exists, and also that

$$\mathcal{P}(x \le X \le x + \delta x) = \Phi(x + \delta x) - \Phi(x)$$

which means in turn that

$$\Phi(x) = \mathcal{P}(X \le x) \tag{9}$$

The cdf has the following properties:

1. $0 \le \Phi(x) \le 1$
2. $\lim_{x \to -\infty} \Phi(x) = 0 \quad \lim_{x \to \infty} \Phi(x) = 1$ or $\Phi(L_b) = 0 \quad \Phi(L_T) = 1$
3. Φ is non-decreasing; i.e. $\Phi(x + h) \ge \Phi(x)$ for $h \ge 0$
4. Φ is right continuous; i.e. $\lim_{h \to 0^+} \Phi(x + h) = \Phi(x)$; 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 sometimes advantages in using the cumulative distribution rather than the pdf. The left panel of Figure 2.1 shows a possible pdf, $\phi(x)$, for a made-up distribution; the right panel shows the corresponding cdf $\Phi(x)$. The dashed lines are the quantiles, which we discuss below.

2.5.1. Lebesgue's Decomposition Theorem

Most treatments of probability take the cumulative distribution function for X as being the more fundamental description of an rv, using equation (9) for the relation to probability. They then define the pdf $\phi(x)$ as the derivative of it, if this exists. The reason for this approach is to allow discrete as well as continuous random variables, though **Lebesgue's 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 $a_i > 0$, 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), and Φ_3 is singular. We can ignore Φ_3 as pathological. Φ_2 has the form $\Phi(x) = \sum_{x_i < x} p_i$, where $p_i = \mathcal{P}(X = x_i)$; that is, the random variable X has a finite probability of occurring at the discrete values x_1, x_2, x_3, …, and zero probability of having any other values. We would then call p_i the **probability mass function** or the frequency function of the random variable X; we avoid the term frequency function because of (later) possible confusion with frequency in the Fourier sense. Distribution functions can thus be applied both for continuous random variables, and for discrete ones—or indeed to both combined (though this combination is pretty unusual). Dice-throwing has been our standard example for a discrete rv; we could also use this to model the probability of finding the geomagnetic field in a normal or reversed polarity state, by assigning integer values to each state, X, say 1 for normal, and –1 to reversed. A less contrived case would be when X applies to the number of some kind of event (for example, number of magnitude 6 or larger earthquakes in a year); this has to be integer-valued, and therefore has to have a discrete distribution.

For such a discrete distribution, the distribution function includes a Φ_2 part; that is, it has steps. While the derivative does not, strictly speaking, exist at these steps, we can obtain the distribution function from the pdf if the pdf $\phi(x)$ contains δ-functions, making ϕ a generalized function. While this approach is mathematically consistent, it is not the one usually followed in probability theory, perhaps because the standard mathematical development of that theory predates the development of generalized functions. This is part of the reason for the preference, in probability texts, for the distribution rather than the density

function: for a discrete random variable, Φ exists and ϕ does not, at least as functions.

2.6. From rv's to Conventional Variables: Means, Variances, Expectations, Moments

While it takes a function to describe a random variable, we often want to summarize certain attributes, such as the "typical" value, or the spread or variability, of a random variable. These attributes are, it must be noted, conventional variables and not rv's: they involve some kind of operation on the pdf (usually integration) the result of which is a conventional variable. In this section we describe some of the operations that can be performed on a pdf to extract summary variables about it: of course this involves considerable compression, so that we lose information but perhaps gain manageability. It is also important to keep in mind that the operations we describe are performed on the pdf's that describe an rv, *not* on data— though these operations on pdf's will certainly suggest how we might form similar summaries on datasets.

One way to "summarize" the central value of a pdf would be to take the value of x at which it attains its maximum value; this is called the **mode**. But density functions may be unimodal (one peak) or multimodal; we showed a multimodal case in Figure 2.1. Given the sensitivity of the mode to the details of the peak of the distribution, it is not a very good measure of the central value.

A better method comes from what are called **moments** of the pdf (a usage that stems from mechanics—remember moment of inertia). To start with, consider a discrete random variable X, with probability distribution $p_i = \mathcal{P}(X = x_i)$, $i = 1, 2\ldots$. (That is, keeping in mind our approach to pdf's through delta functions, the pdf would be $\sum_i p_i \delta(x - i)$). Then we can define the **mean** μ and **variance** σ^2 of the pdf by

$$\mu = \sum_i p_i x_i$$

$$\sigma^2 = \sum_i p_i (x_i - \mu)^2$$

If X is a continuous random variable with pdf $\phi(x)$ then we replace the summations by integrals over the density function (note that we will say "probability distribution" even when we refer to the density function). Then the mean is

$$\mu = \int_{-\infty}^{\infty} x\phi(x)dx \tag{10}$$

and the variance is

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 \phi(x)dx \tag{11}$$

which again measure the central value and spread. In the particular case of the normal distribution, $\sigma = \sqrt{\sigma^2}$ is called the **standard deviation**, but it is probably best not to use this term except when that particular distribution is being used or assumed.

The variance of an rv Y is denoted as $\mathcal{V}[Y]$.

Of course, μ and σ may not completely describe a distribution (though for some distributions they do). Additional characteristics of a probability distribution are given by higher-order moments of the density function. These are defined in two ways: the r-th moments about the origin are, for $r = 1, 2, \ldots$,

$$\mu_r' = \sum_i x_i^r p_i \qquad \text{for } X \text{ discrete}$$

$$\mu_r' = \int_{-\infty}^{\infty} x^r \phi(x) dx \qquad \text{for } X \text{ continuous} \tag{12}$$

and μ_r is the r-th moment about the mean, $r = 2, \ldots$; for X continuous,

$$\mu_r = \int_{-\infty}^{\infty} (x - \mu_1')^r \phi(x) dx$$

which is to say that the mean μ is μ_1'; $\mu_1 = 0$; $\mu_2 = \sigma^2$; and $\mu_2' - \mu^2 = \sigma^2$. The moments of order higher than two give us additional information. The third moment, μ_3, is a measure of the asymmetry or **skewness** of the density function. The next moment, μ_4, is known as the **flatness** (or **kurtosis**).

While the first two moments of a distribution are the commonest measures of central value and spread, there are others, which can be defined on the pdf but become most useful when applied to data with a non-Gaussian distribution. Many of these measures are based on the *quantiles* of the cdf, $\Phi(x)$. Since the cdf increases over its whole range, it has to have an inverse, Φ^{-1}, so we can write

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

Then the p-th quantile of the distribution Φ is the value x_p such that $\Phi(x_p) = p$; equivalently, $\mathcal{P}(X \le x_p) = p$. Thus $x_p = \Phi^{-1}(p)$. The **median** of the distribution, $\tilde{\mu}$, is the quantile corresponding to $p = \frac{1}{2}$, which means that

$$\int_{-\infty}^{\tilde{\mu}} \phi(x) dx = \int_{\tilde{\mu}}^{\infty} \phi(x) dx = \tfrac{1}{2}$$

or

$$\mathcal{P}(X \le \tilde{\mu}) = \mathcal{P}(X \ge \tilde{\mu}) = \tfrac{1}{2}$$

The lower and upper quartiles of Φ, that is to say the quantiles corresponding to $p = 0.25$ and $p = 0.75$, are frequently used as a measure of spread; the difference $x_{0.75} - x_{0.25}$ is known as the **interquartile range**. The dashed lines in Figure 2.1 show how the median and interquartile range are found for a particular cdf; since the pdf has two separate peaks (it is **multimodal**), neither the mean nor the mode are good summaries: actually, in this case it is not clear that the pdf can be well summarized with only one or two numbers. A reason for preferring the median to the mean and variance is that the latter can be heavily influenced by the tails of the pdf; obviously, these have little effect on the median and interquartile range. This kind of summary value that is insensitive to small changes in the pdf is called **robust**.

A more robust measure of spread than the variance is the **mean deviation**, $\tilde{\sigma}$, defined by

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

If we compare this to the definition for the variance, we see that the mean deviation multiplies the pdf by a pair of straight lines to get the function to be integrated, while the variance multiplies by a parabola. Clearly the parabola gives more weight, or influence, to the values of the pdf far from the center—which, as we noted above, may not be a good idea. Both the parabola and the straight lines are examples of what are called **influence functions**; when we discuss robust methods, we will see that a proper choice of such functions can make a big difference in how resistant our estimates are to small changes in the pdf.

2.6.1. Expectations

In the previous section we have shown a number of cases of producing conventional variables from rv's by taking integrals of pdf's multiplied by other functions. This can be generalized to something that is called the **expectation** of a random variable, or of a function of it.

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

$$\int_{L_b}^{L_t} g(x)\phi(x)dx \tag{13}$$

where the limits are those applicable to $g(X)$; for example, if X could range over the entire real line, and $g(x)$ was x^2, the limits for g, and the integration, would be from zero to infinity.

The expectation of an rv Y is denoted as $\mathcal{E}[Y]$.

$\mathcal{E}$ is a kind of operator, like differentiation or integration, taking any random variable (in the form of its pdf) and creating a conventional variable out of it. We say that for any conventional variable c, $\mathcal{E}[c] = c$. Because $\mathcal{E}$ involves integration, it is linear, so that

$$\mathcal{E}[\sum_{i=1}^{k} c_i g_i(X)] = \sum_{i=1}^{k} c_i \mathcal{E}[g_i(X)] \tag{14}$$

The simplest case is when $g(Y)$ is just the variable itself, then we have

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

by the definition of the mean: so the mean is just $\mathcal{E}[X]$. Similarly, equations (10), (11), and (12) become

$$\mathcal{E}[X] = \mu \qquad \mathcal{V}[X] =_{\text{def}} \mathcal{E}[(X-\mu)^2] = \sigma^2 \qquad \mathcal{E}[X^r] = \mu'_r \tag{15}$$

where we use $=_{\text{def}}$ as a shorthand for "the left side is defined to be what is on the right

side".

2.7. Transformations and Functions of Random Variables

We start to look at what might be called the arithmetic of random variables: that is, the rules that give us the pdf for a random variable Y that is in some way related to a random variable X, whose pdf, $\phi_X(x)$, we know. We will see in the next section that something as simple as the sum of two rv's requires a fairly complicated process to produce the pdf of the sum; in this section we deal with the simpler cases of combining rv's with conventional variables, and of functions of an rv.

The most general combination of an rv with conventional variables is a linear transformation, which involves both multiplication and addition; we define a new rv Y as

$$Y = c(X + l) \tag{16}$$

using the variables c and l because we will use them in the next chapter for the spread and location parameters of a pdf. Now consider the probability

$$\mathcal{P}(y \le Y \le y + g) = \int_{y}^{y+g} \phi_Y(v)\, dv \tag{17}$$

From (16) we have that

$$\mathcal{P}(y \le Y \le y + g) = \mathcal{P}(y \le c(X + l) \le y + g) \tag{18}$$

remembering that y is being a limit, is just a conventional variable, and does not change when the random variable does. We can rewrite the right-hand side of (18) as

$$\mathcal{P}\left(\frac{x - l}{c} \le X \le \frac{x - l + g}{c}\right) = \int_{\frac{x-l}{c}}^{\frac{x-l+g}{c}} \phi_X(u)du \tag{19}$$

by the definition of the pdf ϕ_X. In order to make the limits on the integral the same as those in (17), we have to perform a change of variables, with $w = cu + l$ so $u = (w - l)/c$. Making this change, the integral in (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 in the next chapter.

Intuitive though it may be to use the pdf, this is one of many cases in which the cumulative distribution function makes the proof simpler. We can put most of the steps on one line:

$$\Phi_Y(y) = \mathcal{P}(Y \le y) = \mathcal{P}(cX + l \le y) = \mathcal{P}\left(X \le \frac{y - l}{c}\right) = \Phi_X\left(\frac{y - l}{c}\right) \tag{20}$$

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 $Y = g(X)$; how does $\phi_Y(y)$ relate to $\phi_X(x)$? We can in fact use the same approach, provided that $g(X)$ is monotone and differentiable over the range of X, so that there is an inverse function that satisfies $X = g^{-1}(y)$. Then we can follow the steps in (20) and write

$$\Phi_Y(y) = \mathcal{P}(Y \le y) = \mathcal{P}(g(X) \le y) = \mathcal{P}(X \le g^{-1}(y)) = \Phi_X(g^{-1}(y)) \tag{21}$$

which we differentiate, using the chain rule, to get

$$\phi_Y(y) = \frac{d}{dy}\Phi_Y = \frac{d}{dy}\Phi_X(g^{-1}(y)) = \phi((g^{-1}(y))\left|\frac{d}{dy}(g^{-1}(y))\right| \tag{22}$$

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 (22) might appear to provide a simple formula to apply, it is actually better in practice to start with the steps in (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 (22) because there is no unique inverse; but the steps in (21) become

$$\Phi_Y(y) = \mathcal{P}(Y \le y) = \mathcal{P}(X^2 \le y) = \mathcal{P}(-\sqrt{y} \le X \le \sqrt{y}) = \Phi_x(\sqrt{y}) - \Phi_x(-\sqrt{y})$$

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

2.8. Sums and Products of Independent Random Variables

The above discussion has actually avoided what we normally consider to be basic arithmetic operations, such as adding two variables together. We now turn to this by answering the question, given two rv's X_1 and X_2 with known pdf's, what are the pdf's of $X_1 + X_2$ and X_1/X_2? We go through the derivations here, and then make use of the one for summation to demonstrate the Central Limit Theorem (in the next section); we will use the results of this section extensively in the next chapter for deriving a variety of pdf's.

2.8.1. Summing Two Variables

Our first step in answering these questions is one that appears to complicate the problem; we need to generalizing the concept of a pdf to more than one variable. To do this, we introduce the idea of a **joint probability** for random variables. We already have joint probabilities for sets: the joint probability for set A and set B is $\mathcal{P}(A \cap B)$. 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 we can write the joint probability in terms of a pdf of two variables:

$$\mathcal{P}((x_1 \le X_1 \le x_1 + \delta x_1) \cap (x_2 \le X_2 \le x_2 + \delta x_2)) = \int_{x_1}^{x_1+\delta x_1} \int_{x_2}^{x_2+\delta x_2} \phi(x_1, x_2)\, dx_1 dx_2 \qquad (23)$$

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

$$\Psi(y) = \mathcal{P}(Y \le y) = \mathcal{P}(X_1 + X_2 \le y) = \int_{x_1+x_2 \le y} \phi(x_1, x_2) dx_1 dx_2$$

so that the integral is over the shaded area on the left of Figure 2.2.

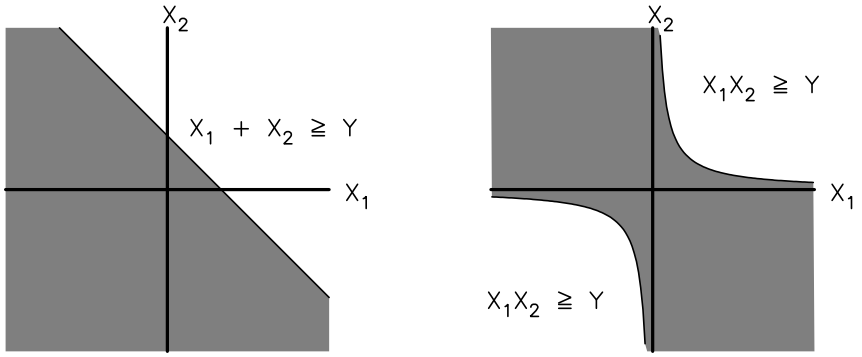

Figure 2.2

We next suppose that X_1 and X_2 are independent; Chapter 4 will deal with the case that they are not. For sets independence means $\mathcal{P}(A \cap B) = \mathcal{P}(A)\mathcal{P}(B)$; this can be consistent with equation (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:

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

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 $X_1, X_2, \ldots, X_n$, 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 will get a random variable whose pdf is the convolution of the component pdf's.

2.8.2. Multiplying Two Variables

For the product of two rv's, we proceed similarly to the derivation for sums: we introduce the rv $Y = X_1 X_2$, with pdf ψ and distribution Ψ; $Y \sim \psi$ with $\psi = \dfrac{d\Psi}{dy}$. Then

$$\Psi(y) = \mathcal{P}(Y \le y) = \mathcal{P}(X_1 X_2 \le y)$$

To get this, we have to integrate the joint pdf $\phi(x_1, x_2)$ over the set such that $x_1 x_2 \le y$; if $x_1 < 0$, $x_2 \ge y/x_1$, while if $x_1 > 0$, $x_2 \le y/x_1$, making the integral of the joint pdf over the shaded area on the right of Figure 2.2. We can write this as the sum of two integrals

$$\int_{-\infty}^{0} \int_{y/x_1}^{\infty} \phi(x_1, x_2) dx_2 dx_1 + \int_{0}^{\infty} \int_{-\infty}^{y/x_1} \phi(x_1, x_2) dx_2 dx_1 \tag{24}$$

We introduce a new variable $s = x_1 x_2$, which makes (24)

$$\int_{-\infty}^{0} \int_{y}^{-\infty} \frac{1}{x_1} \phi(x_1, s/x_1) ds dx_1 + \int_{0}^{\infty} \int_{-\infty}^{y} \frac{1}{x_1} \phi(x_1, s/x_1) ds dx_1 =$$

$$\int_{-\infty}^{0} \int_{-\infty}^{y} \frac{1}{-x_1} \phi(x_1, s/x_1) ds dx_1 + \int_{0}^{\infty} \int_{-\infty}^{y} \frac{1}{x_1} \phi(x_1, s/x_1) ds dx_1 =$$

$$\int_{-\infty}^{y} \int_{-\infty}^{\infty} \frac{1}{|x_1|} \phi(x_1, s/x_1) dx_1 ds$$

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

$$\psi(y) = \int_{-\infty}^{\infty} \frac{1}{|x_1|} \phi(x_1, y/x_1) dx_1 = \int_{-\infty}^{\infty} \frac{1}{|x_1|} \phi_1(x_1) \phi_2\left(\frac{y}{x_1}\right) dx_1 \tag{25}$$

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

$$\psi(y) = \int_{-\infty}^{\infty} |x_1| \, \phi_1(x_1) \phi_2(x_1 y) dx_1 \tag{26}$$

which we will also use in Chapter 3.

2.9. The Central Limit Theorem

In all of probability theory and statistical inference the **normal distribution** (also called the **Gaussian distribution**)[6] plays a major role. The pdf for this, with the mean set to zero, is

[6] For the history of the names, see Stephen S. Stigler (1980). Stigler's Law of Eponymy, *Trans. New York Acad. Sci. Ser. 2* **39**, 147-157 (1980).

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

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

The overriding importance of the normal distribution is 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 distributions of the components.

We now demonstrate this, with two caveats. The first is that this is not a fully rigorous proof. The second is that, despite this theorem, actual data (see, for example, Figure 2.1.1) have a stubborn habit of *not* being normally distributed. Often they are "close enough" that it doesn't matter (much), but the careful data analyst will always check this, and allow for the possibility that the data are non-normal.

2.9.1. The Characteristic Function

We begin by doing something that, unless you are already familiar with convolution, will not be too obvious: we take the Fourier transform of the pdf. We do this because then the convolution operation on the pdf's is replaced by multiplication of their Fourier transforms, which is much more manageable. If we have an rv $X \sim \phi$, we want the Fourier transform of ϕ, which we denote as $\mathcal{F}[\phi]$ or $\hat{\phi}(f)$,[7] and which is given as

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

This is called the **characteristic function** of ϕ; it has the inverse transform

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

Because pdf's are such well-behaved functions (positive and with a finite area) this transform always exists. Note that

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

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

Because of the uniqueness of the inverse Fourier transform (something we have assumed), the characteristic function uniquely determines the pdf, through the inverse Fourier transform, and so completely specifies the properties of X, just as ϕ does. The characteristic function can also be defined in terms of the expectation operator (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

$$\hat{\phi}(f) = \mathcal{E}[e^{-2\pi i f X}] \tag{27}$$

which, mysterious as it might seem at first glance, is just the application of a function to some random variable. Expanding the exponential in equation (27), and making use of the

[7] *Notation alert:* this use of a ^ for the Fourier transform of a function is conventional and convenient. But, you should be aware that we will also use, for example, $\hat{m}$ to denote the estimate of a conventional variable m. In practice the meaning should be clear from the context.

linearity of the expectation operator (equation (14)) and the definition of the higher moments (equation (12)), we find that

$$\hat{\phi}(f) = \sum_{r=0}^{\infty} \frac{(-2\pi i f)^r}{r!} \mu'_r \tag{28}$$

where $\mu'_r = \mathcal{E}[X^r]$. Since a Taylor series (which this is) determines the function uniquely, and the density function is uniquely determined from the characteristic function, we have shown that knowledge of all the moments of a distribution determine it completely.

We can use equation (28) 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 (28), and then evaluate it at zero, we have only one term left in the expansion, so that

$$\hat{\phi}'(0) = -2\pi i \mu'_1 \qquad \text{whence} \qquad \mathcal{E}[X] = = \frac{-\hat{\phi}'(0)}{2\pi i}$$

by (15); similarly, $\hat{\phi}''(0) = -4\pi^2 \mu'_2$, so for the variance we get

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

where, since $\hat{\phi}''(0) < 0$, $\mathcal{V}[X] > 0$ as it should be.

2.9.2. Summing Many Variables

Our actual demonstration of the Central Limit Theorem uses characteristic functions; indeed, the main use of such functions is for proving theorems. To start, we find the characteristic function of the normal pdf, which is

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

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

$$\hat{\phi}(f) = \exp[-2\pi^2 \sigma^2 f^2]$$

Now, suppose we have random variables $X_1, X_2, \ldots, X_n$, which are **independent and identically distributed**, a situation so common that it gets its own acronym, namely **iid**. We assume the pdf has mean of zero, a variance σ^2, and that all the higher moments exist.[8] Let $S_n = \sum_{i=1}^{n} X_i$. The Central Limit Theorem is 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 $\hat{\phi}_n$ is given by

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

Assuming that all the moments of ϕ exist, then so do all the derivatives of $\hat{\phi}$ at $f = 0$ and we

[8] As we will see in the next chapter, there are pdf's for which this is not the case.

can expand $\hat{\phi}$ in a Taylor series:

$$\hat{\phi}(f) = \hat{\phi}(0) + \frac{f}{1!}\hat{\phi}'(0) + \frac{f^2}{2!}\hat{\phi}''(0) + \ldots = 1 + \frac{f^2}{2!}\hat{\phi}''(0) + \frac{f^3}{3!}\hat{\phi}'''(0) + \ldots$$

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

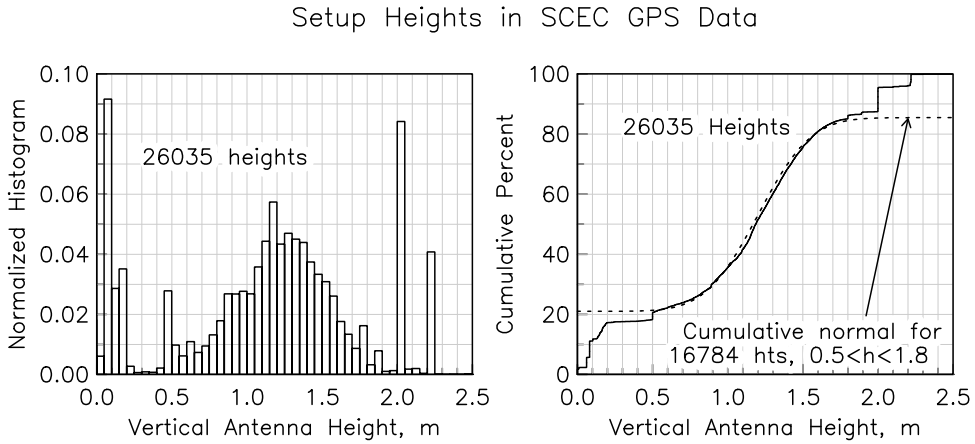

Figure 2.4

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

$$\hat{\phi}_n(f) = \exp\left[n \ln\left(1 + \frac{f^2}{2!}\hat{\phi}''(0) + \frac{f^3}{3!}\hat{\phi}'''(0) + \ldots \right) \right]$$

$$= \exp\left[\frac{nf^2}{2!}\hat{\phi}''(0) + \frac{nf^3}{3!}\hat{\phi}'''(0) + \ldots \right]$$

where we have used the power series expansion $\ln(\varepsilon) = 1 + \varepsilon + \varepsilon^2/2 + \ldots$. Next we define a new variable

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

and a constant

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

Then the series can be rewritten as

$$\exp\left[-2\pi^2\sigma_n^2 f^2 + (\sigma_n f)^3 \frac{c_3}{n^{\frac{1}{2}}} + O\left(\frac{(\sigma_n f)^4}{n} \right) \right]$$

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 0 and variance $n\mathcal{V}[X] = n\sigma^2$, which is what we wanted to show.

To go from mathematics to real data, Figure 2.4 shows the histogram, and the cumulative distribution function 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. In the right frame, the dashed line shows that, indeed, the cumulative distribution function for these heights is very nearly a Gaussian, or Normal, distribution: since these data represent nearly 17000 decisions by hundreds of people over two decades, it is impressive that they can be described so simply. This, and other distributions of random variables, will be the subject of the next chapter.