1 Random Variables and Probability Distributions: Formalizing Uncertainty

Uncertainty is everywhere. Being able to organize and process uncertainty helps distinguish good from mediocre planners. An essential planning task is making predictions in the face of uncertainty: How many children of school age will live in a new development? What peak level of traffic can be expected in a vacation resort? How long will a tenant occupy an apartment? How many graduates of a job-training program will be employed six months after graduation? How likely is it that a small town will employ a professional city manager if it grows to 50,000 people? What is the likelihood that a patient will have to wait more than an hour to be treated in a city hospital emergency room? The list of questions is endless, and the questions lurk in every corner of city planning. This chapter is about a language for answering such questions.

The language is *probability theory*. You will see only the bare bones of the theory in this book, but you should learn enough to acquire useful skills and to be able to learn more as necessary. Probability is about *random variables*, which are variables whose values are not certain ahead of time. The answer to each question above is a random variable. For instance, the number of children in the new development could range anywhere from zero to a large number. In two identical developments the numbers of children are likely to be different, and in any one development the number will change over time. If we could only peek ahead in time, we could actually count the children of the future tenants; but we cannot, so we must make an educated guess. Our education for guessing derives from our knowledge of other similar developments and of typical tenants and perhaps from our own intuitions about the particular development in question.

We express our knowledge about the possible values of a random variable in terms of its *probability distribution*. To every possible value of the random variable there corresponds a number that represents the probability the random variable will take on precisely that value. The collection of possible values and their probabilities is the probability distribution. The probability assigned to each value of the random variable is a fraction between 0 and 1.0. If the probability is zero, the corresponding value can never occur; if the probability is unity, the corresponding value will certainly occur. The full set of possible values of the random variable must be "mutually exclusive and collectively exhaustive," meaning that one and only one of the possible values will

actually occur. Since the random variable must take on some value, the probabilities assigned to all the possible values must sum to 1.0.

As an example, consider the eleven towns in Massachusetts with 1975 populations of 30 to 35,000. Suppose we wish to study their type of local government. The random variable of interest is governmental structure. It can take on three values: open town meeting, representative town meeting, and city council. Note that the random variable need not be a number; in this case it is a category. Note also that the governmental structure is not random in the sense that the type of town government changes from day to day like the weather. Each town among the eleven has one form of government and has doubtless had that particular form for a long time, and if someone named a particular town and asked us its form of government, we would look up the answer. But if the question were put more generally, we would have to use a probability distribution: "What is the chance that any given town has a representative town meeting?" To answer this, we turn to the probability distribution of the random variable shown in table 1.1. The answer is that, not knowing the identity of the town (other than that it is in Massachusetts and has 30 to 35,000 people), there is a 0.37 probability that the town uses a representative town meeting. Our formal notation for expressing this result will be

Prob [town has representative town meeting] = 0.37.

Of course, any particular town either does or does not have such a

Table 1.1
Form of local government in 11 Massachusetts towns of 30 to 35,000 population

Form of government	Number of towns	Fraction of towns
Open town meeting		0.18
Representative town meeting	4	0.37
City council	5	0.45
	11	1.00

Source: Massachusetts League of Cities and Towns, *Municipal Directory*, 1975–1976.

forum, but all we can do when considering an anonymous town is give the relative likelihoods of the alternative forms of government.

We can now make two observations about predictions using our probability distribution. First, if someone asked us, "Which type of government do you think a given town has," we should answer, "City council." This is the modal, or most frequently occuring category, and is our best bet for a guess. We stand a 45 percent chance of being right, compared to only 37 and 18 percent for the other possible guesses. We still have uncertainty, but we have marshalled our evidence and used it to improve on a blind guess which treats all three answers as equally likely. Second, if someone asked us how many of 100 towns of 30 to 35,000 population use either form of town meeting, we have our probability distribution as a starting place. If all the towns were like those in Massachusetts, we would expect about 55 to have town meetings; if all 100 towns were in the midwest, we might expect that 55 would be an upper bound on the number, presuming that town meetings are more common in New England than elsewhere. In this case we are using the data to inform a subjective estimate of probabilities.

It is important that you be able to read and write probability distributions. There are two varieties, corresponding to discrete and continuous random variables. A *discrete random variable* only takes on values from a distinct set: the three types of local government in the example above, the number of children in the new development, the number of users of a neighborhood health center. Probability distributions for discrete random variables all look generally like that shown in figure 1.1a. The height of each vertical line represents the probability that the random variable takes on the value in question. If you stack the lines end to end they must form a line of length 1.0 (the probabilities must sum to 1.0). Note that some probabilities (like the fifth value in the graph) will be zero, meaning that the random variable can never take that value.

Continuous random variables are not so restricted in the values they can take on—there are an infinite number of possible values. An example would be the length of time spent waiting for treatment in a city hospital emergency room. There is no reason to expect this wait to occur exactly in intervals of 5 minutes or 1 minute; although we may ultimately be forced to record the time in discrete units of 1 second intervals, in principle

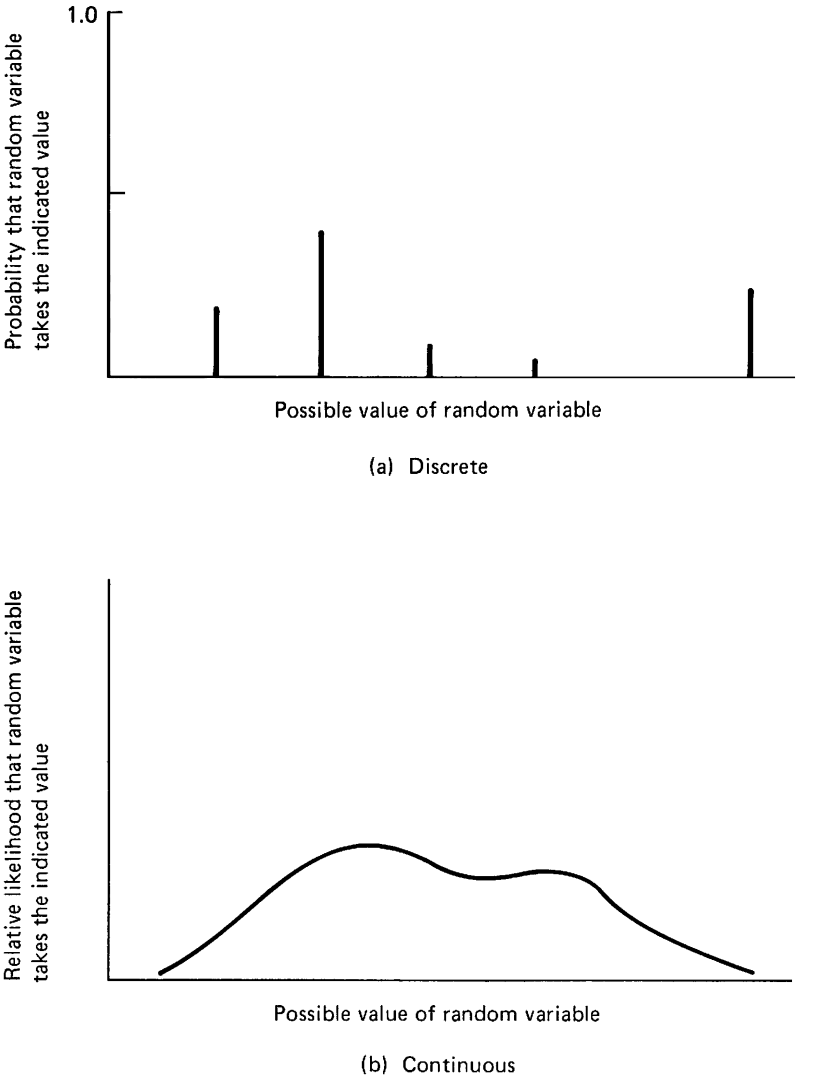

Figure 1.1
Examples of probability distributions for discrete and continuous random variables

we can treat the waiting time as if it were perfectly continuous. In the case of continuous random variables, the probability distribution is a smooth curve as in figure 1.1b, not a sequence of vertical lines, and shows not the actual probability that the random variable takes on the particular value, but the probability relative to other possible values. While the curve cannot drop below the horizontal axis, it need not stay below unity since the height of the curve represents the relative likelihood of the value below it, not the actual probability as in the discrete case. However, just as the sum of the discrete probabilities must equal unity so must the area under the curve of the continuous probability distribution. For continuous random variables we can ask only interval questions, for example, "What is the probability that the waiting time is *between* 15 and 20 minutes"; or "What is the chance that the waiting

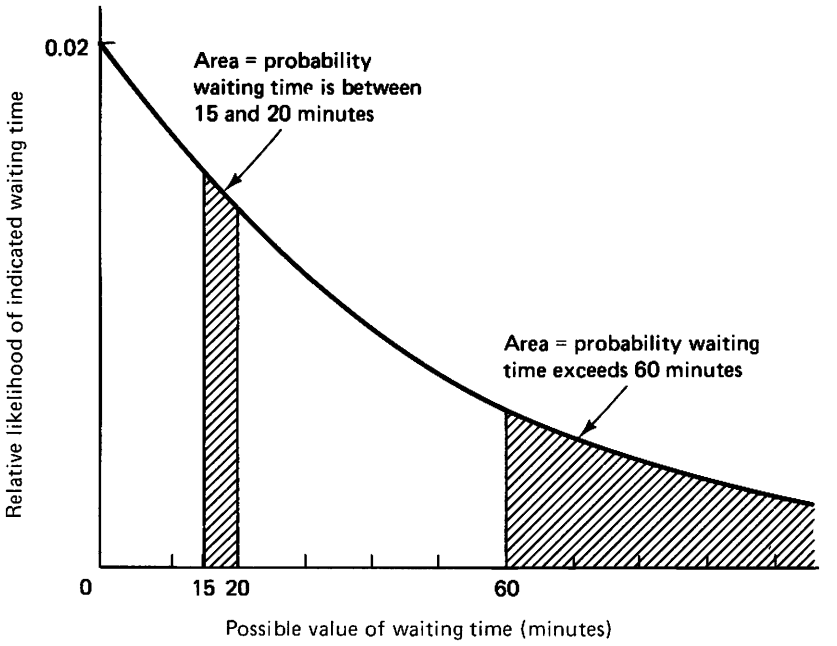

Figure 1.2
Hypothetical probability distribution of waiting time in a city hospital emergency department

time *exceeds* 60 minutes?" The answers to these questions are the corresponding areas under the curve, as shown in figure 1.2 (the entire area must equal 1.0).

Of course, we can ask similar interval questions of discrete random variables. Consider the discrete distribution shown in figure 1.3 for the number of patients who arrive at the emergency room during a certain hour. From figure 1.3 it appears that the probability that seven or more patients will arrive is less than 0.01, the probability that zero or one will arrive is about 0.40, and the probability of three to five arrivals is also about 0.40. These compound probabilities are obtained by adding together the appropriate individual probabilities: the probability of zero or one arrivals is the probability of zero arrivals plus the probability of one arrival.

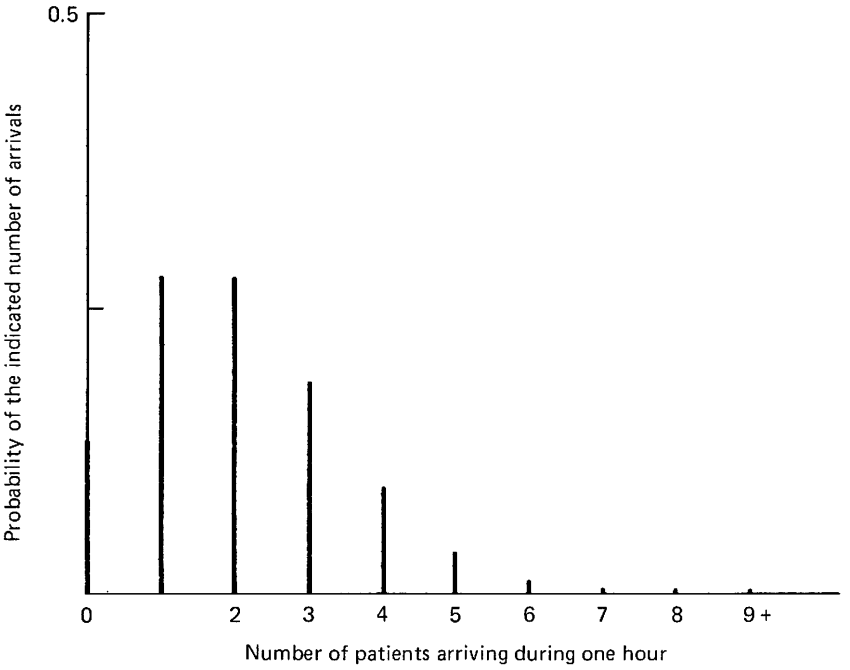

Figure 1.3
Hypothetical probability distribution of the number of patients arriving at a city hospital emergency department during one hour

Where do the probability distributions come from? That in figure 1.3 follows a theoretical pattern (a Poisson distribution with a parameter 2.0, see chapter 4), and the probabilities are tabulated in statistical reference books or can be computed from a formula. In practice, we might not have a fully developed theory, so we might rely heavily on measurements, counting the number of arrivals at the emergency room during the same hour on a number of different days, and construct a *histogram* of the number of days for which the arrivals totaled zero, one, two, and so on. Such a histogram is shown in figure 1.4, which displays the results of 26 (2 + 7 + 9 + 3 + 4 + 0 + 1) observation days. On 3 of the 26 days there were exactly 3 patients arriving during the hour chosen for study; on 9 of the 26 days exactly 2 patients arrived. It is a trivial matter to convert the histogram into a legitimate probability

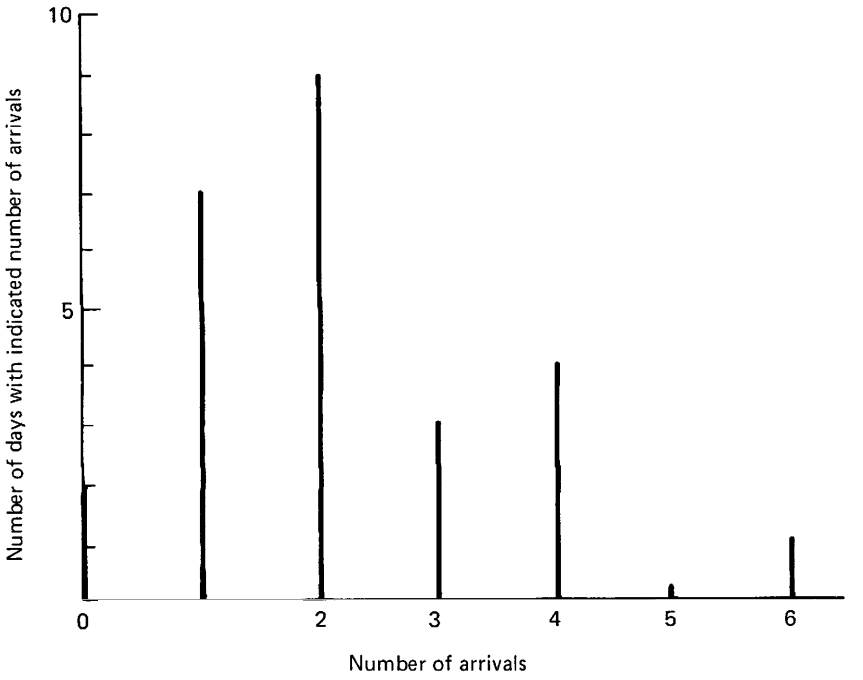

Figure 1.4
Histogram of the number of patients arriving at a city hospital emergency department during one hour

distribution: just divide each number of days by the total (26) to make the corresponding probability. Thus, for instance, the estimate from the 26 days of observations is that there is a $4/26 = 0.15$ probability of exactly 4 arrivals.

You should never fall into the trap of believing that doing statistics is a mindless, automatic process free of substantive judgment; the design of histograms offers a simple but telling instance. I use the word "design" consciously, since a statistical construct like a histogram is as much an artifact as a built structure—having many possible forms, of which one is chosen to fulfill a function. Both the component parts of the histogram and their organization must be selected carefully by the analyst. In the case of the histogram of the number of emergency department arrivals during one hour, the first design choice involves the definition of "arrival." Since some nonemergency cases arrive at the emergency department and are immediately directed to some other part of the hospital (or out of the hospital altogether), there must a decision made as to whether those people who receive no care count as arrivals. Likewise, family members accompanying a patient may count as arrivals for the facility designer who must arrange for their seating but not for the hospital administrator who must determine the physician staffing level. A second design choice involves which time periods to use for the hourly counts of arrivals. If the arrival rate varies significantly by time of day (as it almost always does for urban emergency services), then only counts for the same hour of the day can be used. But if the arrival rate also varies significantly by day of the week, then it may be necessary to gather data only once each week rather than once every day. These decisions about data pooling depend both on the nature of the random process generating the data and the use to which the histogram will be put. A third design choice involves the selection of categories for display of the data. Here the histogram designer feels two opposing pressures. To preserve detail, the designer wishes a large number of categories, yet to preserve compactness and smoothness in the display he wishes a small number of categories. This tension was resolved in figure 1.3 by combining all large numbers of arrivals into the category 9+. The resolution is usually less obvious when the random variable in question is continuous and has no natural categories, as in the case of waiting times in the emergency department. Finally, it may happen that there are no data available

on which to base a histogram, or the available data are not directly applicable (perhaps arising from the wrong time or place), so the distribution becomes an exposition of the planner's best subjective judgment. In all cases, the histogram is the planner's creation, formalizing his uncertainty.

Summary

Much of planning involves a confrontation with uncertainty—predicting the values to be taken on by random variables. By a combination of empirical and / or subjective methods we summarize in a probability distribution our knowledge of the relative likelihoods of the various possible values of a random variable. The form of the distribution varies depending on whether the random variable is discrete or continuous, but the knowledge provided is always the probability that the random variable takes on some particular value or set of values.

References and Readings

Davis, K. "World Urbanization 1950–70." In L. S. Bourne and J. W. Simmons, ed., *Systems of Cities: Readings on Structure, Growth and Policy*. New York: Oxford University Press, 1978.

Tukey. "Scratching Down Numbers," chapter 1.

Problems

1.1

The table that follows reports on ambulance response time (the time delay between calling an ambulance and its arrival at the scene of an emergency) in rural areas around Wheeling, W. Va. Use the table to plot a histogram of response time. Note that the table itself reports cumulative response time.

Time (minutes)	Percentage of calls answered
5	24
10	62
15	81
20	89
25	93
30	96
>30	100

1.2

The following data show the number of new housing units authorized by building permits in Belmont, Mass., from 1960 to 1974. Prepare a histogram of the number of units authorized each year.

Year	Number of units authorized
1960	55
1961	67
1962	108
1963	81
1964	66
1965	37
1966	44
1967	14
1968	104
1969	12
1970	364
1971	25
1972	30
1973	27
1974	20

1.3

The following data show the number of home sales in Boston neighborhoods over a 2-year period. Use these data to make a histogram illustrating the change in level of sales from 1975–76 to 1976–77. Justify the choices you make regarding the issues of representing the changes in each neighborhood (absolute vs. percentage differences) and aggregating the cases into groups.

	Home sales	
Neighborhood	7/75 to 6/76	7/76 to 6/77
Roxbury	266	148
North Dorchester	663	479
South End	180	144
Jamaica Plain	258	200
South Boston	250	176
West End	134	84
South Dorchester	686	536
Charlestown	154	83
East Boston	275	200
Roslindale	292	218
Back Bay–Fenway	361	325
Hyde Park	300	318
Allston–Brighton	297	241
North End	45	53
West Rosbury	292	233