CONTENTS

CHAPTER

1

SAMPLING THEORY

> *"Very well," said Henchard quickly, "please yourself. But I tell you, young man, if this holds good for the bulk, as it has done for the sample, you have saved my credit, stranger though you be. What shall I pay you for this knowledge."*
>
> THOMAS HARDY—*The Mayor of Casterbridge*

PROBABILITY DISTRIBUTIONS

Probability is a character associated with an event indicating its tendency to take place. To speak of probabilities one needs to define an *event-space* that contains all possible (or known) outcomes of a certain process. If there are n possible outcomes associated with a certain process and an event, A, occurs in m of these outcomes, then the probability of obtaining event A, $P(A)$, is defined as

$$P(A)=\frac{m}{n}. \tag{1.1}$$

This definition has the convenience of expressing probabilities as real numbers between 0 and 1, inclusive. Probabilities calculated as defined above are also known as *relative* or *objective probabilities*.

In many situations it is difficult, if not impossible, to enumerate all possible or known events. For example, it is hard to define an event-space to adequately calculate the probability that oil prices will change or that the University of Washington Huskies will win the Rose Bowl. The probabilities of these and similar events can be estimated from experience and trends. Such probabilities are known as *subjective probabilities*. Our discussion will be restricted to objective probabilities.

A plot of events versus their probabilities constitutes a probability

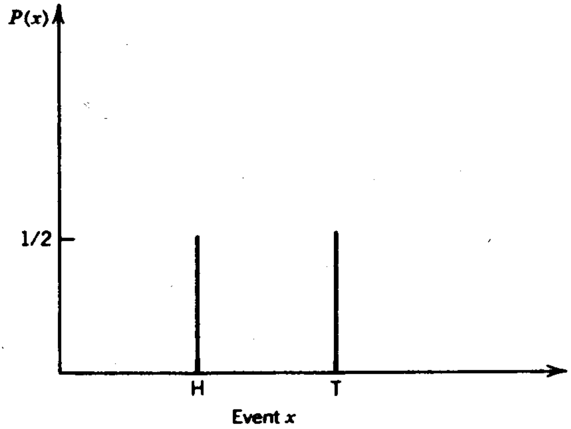

Figure 1.1 Probability distribution of tossing a fair coin.

distribution. The probability distributions of tossing a fair coin and rolling a fair die are shown in Figures 1.1 and 1.2, respectively. In each of these examples all events have the same probability. This is not generally the case. Consider the probability distribution of rolling a pair of dice with the outcome being taken as the sum of the digits on each die. The probability distribution of such a process is shown in Figure 1.3. For

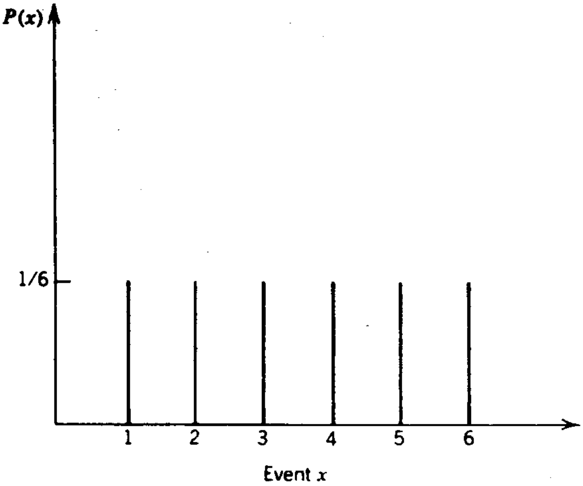

Figure 1.2 Probability distribution of rolling a fair die.

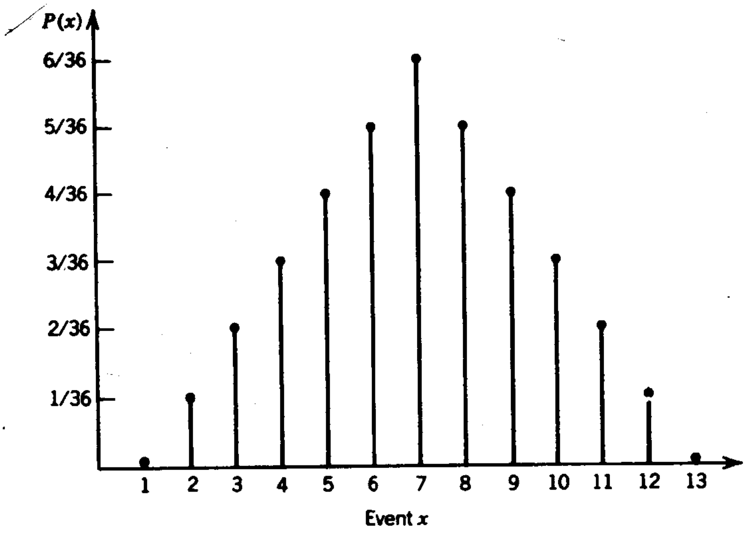

Figure 1.3 Probability distribution of rolling a pair of dice.

convenience, the event-space may be considered continuous. If it is assumed that all values between 1 and 13 are possible, and that extrapolation is valid, Figure 1.3 can be redrawn as shown in Figure 1.4.

Assume that the heights of the students at a certain school are being measured. The measurements may range from 150 to 200 cm. Even though our measurement may be precise only to 0.5 or 1.0 cm, we can assume that our event-space (the height) is continuous with any value between 150 and 200 cm possible. The probability distribution of such an event may look like that shown in Figure 1.5.

Probability distributions can assume any shape depending on the event-space under consideration. Distributions that show variations in probability (usually maximizing at a particular event) are of great importance in chemical analysis. Analysts rarely deal with probability distributions of the type shown in Figures 1.1 and 1.2. For convenience, a probability distribution is expressed in terms of a ***probability density function***. Generally, if

$$P(x_a < x < x_b) = \frac{\int_{x_a}^{x_b} f(x)\,dx}{\int_{-\infty}^{+\infty} f(x)\,dx}, \tag{1.2}$$

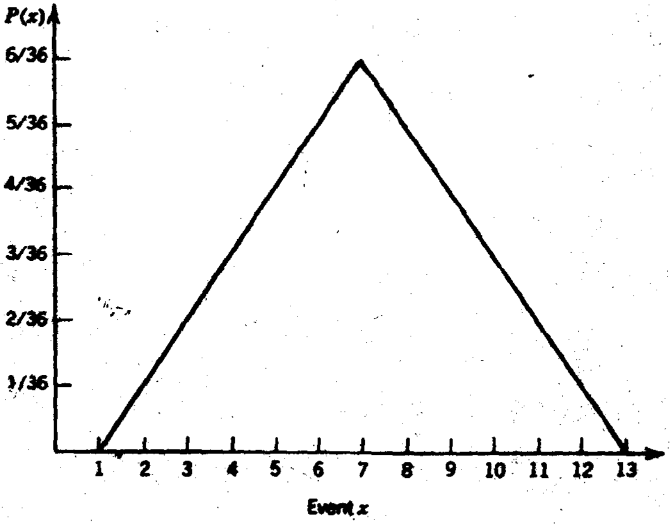

Figure 1.4 Figure 1.3 if the event-space is considered continuous.

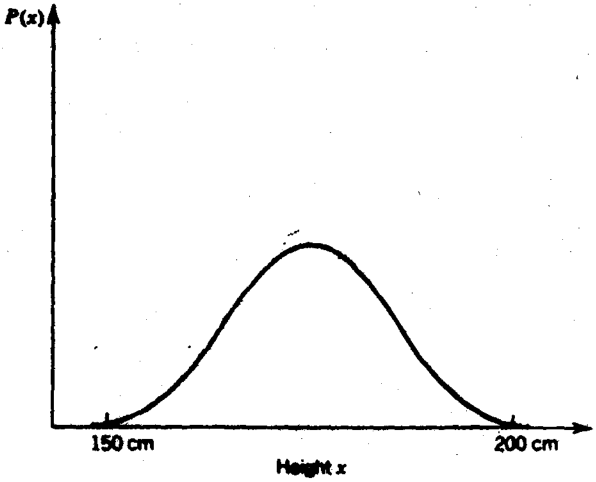

Figure 1.5 Probability distribution of the heights of students in a school.

then $f(x)$ is called the probability density function for the variable x. If the probability distribution shown in Figure 1.5 is expressed in terms of a probability density function, $f(x)$, then the probability of finding a student whose height is between 185 and 190 cm is the ratio of the shaded area to the total area under the curve in Figure 1.6.

The most commonly studied probability distribution is the normal (also called the Gaussian) distribution (1). For any normally distributed variable, x, the probability density function, $f(x)$, is given by

$$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \qquad -\infty<x<\infty. \tag{1.3}$$

Thus the normal distribution is defined only by two parameters. The first is the mean, μ, and the second is the variance, σ^2. These parameters are given by

$$\mu=\int_{-\infty}^{\infty} xf(x)\,dx \quad \text{or} \quad \mu=\frac{\sum_{i=1}^{N} x_i}{N} \tag{1.4}$$

and

$$\sigma^2=\int_{-\infty}^{\infty}(x-\mu)^2 f(x)\,dx \quad \text{or} \quad \sigma^2=\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2, \tag{1.5}$$

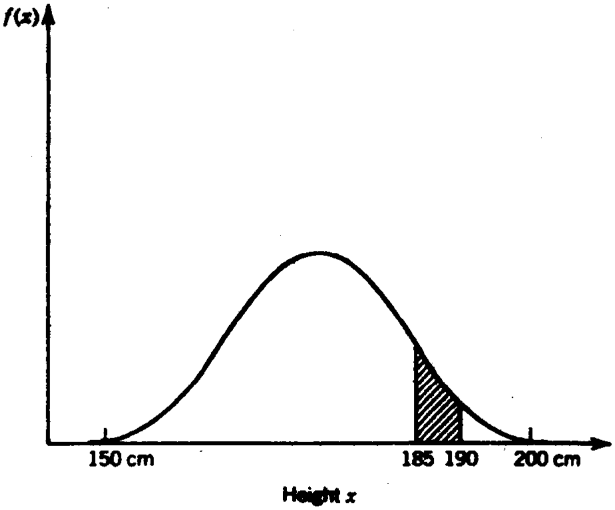

Figure 1.6 Probability density function associated with Figure 1.5. Probability of finding a student whose height is between 185 and 190 cm is the ratio of the shaded area to the total area.

where N is the total number of elements under consideration. A normal distribution of a variate x is expressed as $x = N(\mu, \sigma^2)$ and has the following properties:

1. It has a maximum at $x = \mu$.
2. It is symmetric with respect to $x = \mu$.
3. It has two points of inflection at $x = \mu \pm \sigma$.
4. A change in μ causes a translation of the curve without changing its shape.
5. A change in σ^2 will widen or narrow the curve without a change in μ.

STANDARD NORMAL VARIATE

Consider a variate $x = N(\mu, \sigma^2)$. Define Z such that

$$Z_i = \frac{x_i - \mu}{\sigma}. \tag{1.6}$$

The mean of the Z_i's, $\bar{Z}$, is

$$\bar{Z} = \frac{\sum_{i=1}^{N} Z_i}{N} = \frac{1}{N\sigma}\left[\sum_{i=1}^{N} x_i - \sum^{N} \mu\right]$$

$$= \frac{1}{N}[N\mu - N\mu] = 0.$$

The variance of the Z_i's, σ_Z^2, is

$$\sigma_Z^2 = \frac{\sum^{N} (Z_i - \bar{Z})^2}{N} = \frac{\sum^{N} (Z_i)^2}{N}$$

$$= \frac{1}{N} \frac{\sum^{N} (x_i - \mu)^2}{\sigma^2} = \frac{\sigma^2}{\sigma^2} = 1.$$

Therefore $Z = N(0, 1)$.

The probability density function is given by

$$f(Z) = \frac{1}{\sqrt{2\pi}} e^{-Z^2/2}.$$

This function is also known as the ***standard normal function***. It describes all normally distributed variates regardless of the different values of their parameters (i.e., μ and σ^2). Therefore, all normal distributions with different means and variances, after transformation to a standard form by Equation (1.6), can be represented by a single table of probabilities. The probabilities of the standard normal variable Z are given in Table A.1 in the appendix.

POPULATIONS AND SAMPLES

When a phenomenon like the heights of students is to be studied, two statistical approaches are usually used. The first is to measure the height of every student of the N students enrolled in the school. Conducted in this manner one can find the "true" average height (i.e., μ). The data set of such an experiment contains information about every single element under observation and is thus called ***population data***. Parameters such as μ and σ^2 calculated from such data are referred to as ***population parameters***. The values of μ and σ^2 are calculated according to Equations (1.7) and (1.8):

$$\mu = \sum_{i=1}^{N} \frac{x_i}{N} \tag{1.7}$$

and

$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}. \tag{1.8}$$

The second approach is to measure the heights of only a group (***sample***) of n students. This approach may be advantageous in terms of economics and time. One can calculate an average, $\bar{x}$, and a variance, s^2 (***sample parameters***) from the sample data. The sample parameters are calculated according to Equations (1.9) and (1.10):

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} \tag{1.9}$$

and

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}. \tag{1.10}$$

The crucial question at this point is: How do $\bar{x}$ and s^2 differ from the true values μ and σ^2, respectively? Before an adequate answer to this question is considered, the sampling process must be examined.

The chosen sample must be *representative* of the population. This can be guaranteed if we have a *random sample*; that is, when every element in the population has an equal chance of being included in the sample. A random sample can be obtained, for example, by drawing names from a well mixed box that contains the names of all the students. Our sample will not be random if, for instance, we choose the students from The Dean's Honor List (indeed, the authors would not be chosen). Sample size is also of great importance. The larger the sample size the closer the agreements between σ^2 and s^2 and between μ and $\bar{x}$. In practice, sample size is usually determined by the economics of the experiment being conducted.

As depicted in Figure 1.7, the sample distribution can be anywhere within the parent distribution. The calculated sample mean can be a good approximation of μ (e.g., $\bar{x}_B$) or a bad one (e.g., $\bar{x}_A$ or $\bar{x}_C$). Samples A and C may be the product of nonrandom sampling or a small sample size.

The mean, $\bar{x}$, and the variance, s^2, describe the distribution of one sample ($x \pm rs$, where r is a real number). Also of interest is the *distribution of the means*. If m different samples are obtained and m different averages are calculated, one can treat the m means as a separate population. An average, $\bar{m}$ or $E(\bar{x})$, and a variance, σ^2_m or $\sigma^2_{\bar{x}}$, can be obtained for this distribution. These important parameters can be

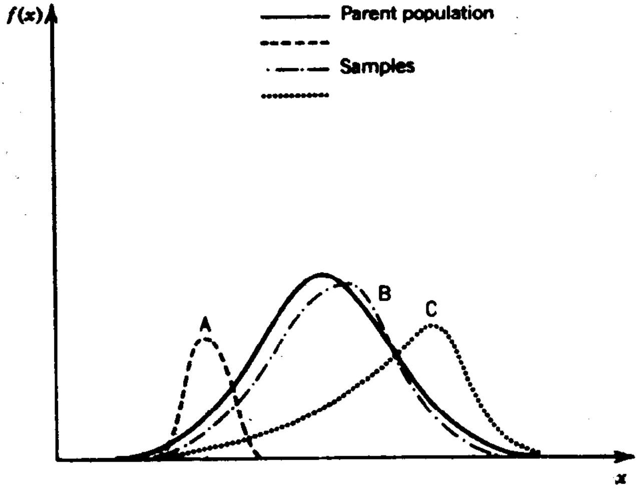

Figure 1.7 The sample distribution can be anywhere within the parent distribution.

estimated from one normally distributed random sample. The sample average, $\bar{x}$, is an unbiased estimate of $\bar{m}$:

$$\bar{m} = E(\bar{x}) = \frac{\sum_{i=1}^{n} x_i}{n}. \tag{1.11}$$

The variance of the mean, $\sigma_{\bar{x}}^2$, can be derived as follows. From Equation (1.9),

$$\bar{x} = \frac{x_1}{n} + \frac{x_2}{n} + \cdots + \frac{x_n}{n}.$$

Let Var(x) denote the variance of variate x. By using the propagation of error principle (2) and assuming that the x_i's are uncorrelated:

$$\mathrm{Var}(\bar{x}) = \left(\frac{1}{n}\right)^2 [\mathrm{Var}(x_1) + \mathrm{Var}(x_2) + \cdots + \mathrm{Var}(x_n)].$$

Assuming that the x_i's have equal variances, s^2, the above equation can be rewritten:

$$\mathrm{Var}(\bar{x}) = \left(\frac{1}{n}\right)^2 ns^2 \quad \text{or} \quad s_{\bar{x}}^2 = \frac{s^2}{n}. \tag{1.12}$$

Thus the average and the variance of the normally distributed means can be estimated from Equations (1.11) and (1.12). When these parameters are known, the interval in which $\bar{x}$, the mean of any random sample drawn from the same population, exists can be estimated. Table A.1 shows that the value of a normally distributed variate lies in the interval $\bar{x} \pm 1.96s$ 95% of the time. By the same token, the mean of the population, μ, as calculated from any random sample taken from the population, lies in the interval $\bar{x} \pm 1.96s_{\bar{x}}$ 95% of the time. Generally, for a random sample of size n, the range

$$\bar{x} - Z\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + Z\frac{s}{\sqrt{n}} \tag{1.13}$$

is an estimate of μ at a certain confidence level defined by Z. Some popular values of Z are given in Table 1.1. Thus a random sample of size n, drawn from a population of size N, has an average $\bar{x}$ and a variance s^2. The probability that the true mean, μ, is in the interval $\bar{x} \pm 1.96s/\sqrt{n}$ is 0.95 (95% of the time). The variance of the mean, $s_{\bar{x}}^2$, may be corrected for the finite sample size and expressed as

$$s_{\bar{x}}^2 = \left(1 - \frac{n}{N}\right)\frac{s^2}{n}. \tag{1.14}$$

Table 1.1 Values of Z at Various Confidence Levels

Z	Confidence Level
1.64	90%
1.96	95%
2.58	99%

The sample variance is an unbiased estimator of the variance of the population,

$$\sigma^2 = s^2. \tag{1.15}$$

If the population is not normally distributed, large random samples ($n > 30$) must be taken to estimate μ and σ^2 (3). The means of the random samples, however, are considered normally distributed (3, 4). The reliability of the estimates of μ and σ^2 ($\bar{x}$ and s^2) requires that s^2 be a good approximation of σ^2. Although large samples are needed, this sampling technique has the advantage of tolerating non-normally distributed parent populations. If the parent population is known to be normal, only a small random sample is sufficient to estimate μ and σ^2. However, different statistics are needed to treat small samples taken from normally distributed parent populations.

STUDENT'S *t*-DISTRIBUTION

Take all possible small random samples of size n from a normally distributed population with a mean μ and variance σ^2. For each sample compute $\bar{x}$ and s^2 and define

$$t = \left(\frac{\bar{x} - \mu}{s}\right)\sqrt{n}. \tag{1.16}$$

The distribution of the t values is known as the Student's t-distribution (5). Every value of n gives rise to a characteristic t-distribution that will be associated with $n - 1$ degrees of freedom. The variable t exists within a certain range given a certain sample size and a particular confidence level. In general,

$$\bar{x} - t\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t\frac{s}{\sqrt{n}}, \tag{1.17}$$

where the value of t is defined by the confidence level and $(n-1)$. If $n=6$ and the desired confidence level equals 0.95, Table A.2 in the appendix shows t to be 2.57 for 5, or $(n-1)$ degrees of freedom. Therefore, the 0.95 confidence level interval for μ is given by

$$\bar{x}-2.57\frac{s}{\sqrt{6}}\leq\mu\leq\bar{x}+2.57\frac{s}{\sqrt{6}}. \qquad (1.18)$$

Equation (1.18) implies that μ exists 95% of the time in the range $\bar{x}\pm 2.57s/\sqrt{6}$. To estimate σ^2 we refer to yet another distribution—the χ^2 distribution. We define

$$\chi^2_{n-1}=\frac{(n-1)s^2}{\sigma^2}. \qquad (1.19)$$

The population variance, σ^2, can be estimated from the sample variance, s^2. An interval in which σ^2 exists is given by

$$\frac{(n-1)s^2}{\chi^2_{(1-\alpha/2,n-1)}}\leq\sigma^2\leq\frac{(n-1)s^2}{\chi^2_{(\alpha/2,n-1)}}, \qquad (1.20)$$

where $1-\alpha$ is the desired confidence level. Assume that ten determina'ions for lead in Lake Sammamish are reported with a standard deviation, s, of 4 mg/L. From Table A.3 in the appendix, the population variance, at the 98% $(\alpha=.02)$ confidence level, is in the following interval:

$$\frac{9(16)}{\chi^2_{(.99,9)}}\leq\sigma^2\leq\frac{9(16)}{\chi^2_{(.01,9)}}$$

$$6.65\leq\sigma^2\leq 68.90.$$

To summarize, one only needs small samples if the population is known to be normally distributed. Equation (1.17) calculates an *exact* confidence level interval for μ. If the population is not normally distributed, we have to obtain large samples $(n>30)$ and Equation (1.13) gives an *approximate* confidence interval for μ. Sample size and the allowable error (the degree of accuracy required) are important since their values determine the estimation interval. If we denote the maximum allowable error by ε and the variance in the population whose mean is being estimated by σ^2, the sample size needed, n, is given by

$$n=\frac{Z^2\sigma^2}{\varepsilon^2}, \qquad (1.21)$$

where Z is the value of the standard normal variate associated with the desired confidence level. We can obtain an estimate of σ^2 from past experience or from a preliminary sample. Alternately, one can manipulate sample size, n, until

$$\varepsilon = Zs_{\bar{x}}. \tag{1.22}$$

Thus far, we have considered parameter estimation of population distribution from only one sample. Better estimations are possible if more than one random sample are drawn from the population. If all possible samples of size n are drawn from a population of N, one can calculate $\bar{x}$ for each sample. The $\bar{x}$'s will be normally distributed when $n > 30$ even if the parent population is not. If the parent population is normally distributed, the $\bar{x}$'s will be normally distributed regardless of n. The distribution of $\bar{x}$'s will have a mean, $E(\bar{x})$, and a variance, $s^2_{\bar{x}}$. Population parameters are given by

$$\mu = E(\bar{x}) \tag{1.23}$$

and

$$\sigma^2 = s^2_{\bar{x}} \frac{n(N-1)}{(N-n)}. \tag{1.24}$$

If $n \ll N$,

$$\sigma^2 = ns^2_{\bar{x}}. \tag{1.25}$$

BINOMIAL DISTRIBUTION

Consider a particular outcome, O, of a single trial of a certain process and let

$p =$ probability that O will take place

and

$q =$ probability that O will not take place

such that $p + q = 1$.

The binomial distribution is a discrete distribution that predicts the probability of O taking place X times in N trials. This is given by

$$P(X, N) = \frac{N!}{X!(N-X)!} p^X q^{N-X}. \tag{1.26}$$

Consider the tossing of a fair coin. In a single trial the probability of getting a head, p, is $\frac{1}{2}$. The probability of *not* getting a head is equal to