1

A review of basic probability theory

This is a book about the applications of probability. It is hoped to convey that this subject is both a fascinating and important one. The examples are drawn mainly from the biological sciences but some originate in the engineering, physical, social and statistical sciences. Furthermore, the techniques are not limited to any one area.

The reader is assumed to be familiar with the elements of probability or to be studying it concomitantly. In this chapter we will briefly review some of this basic material. This will establish notation and provide a convenient reference place for some formulas and theorems which are needed later at various points.

1.1 PROBABILITY AND RANDOM VARIABLES

When an experiment is performed whose outcome is uncertain, the collection of possible **elementary outcomes** is called a **sample space**, often denoted by Ω. Points in Ω, denoted in the discrete case by ω_i, $i = 1, 2, \ldots$ have an associated probability $P\{\omega_i\}$. This enables the probability of any subset A of Ω, called an **event**, to be ascertained by finding the total probability associated with all the points in the given subset:

$$P\{A\} = \sum_{\omega_i \in A} P\{\omega_i\}$$

We always have

$$0 \leqslant P\{A\} \leqslant 1,$$

and in particular $P\{\Omega\} = 1$ and $P\{\varnothing\} = 0$, where $\varnothing$ is the empty set relative to Ω.

A **random variable** is a real-valued function defined on the elements of a sample space. Roughly speaking it is an observable which takes on numerical values with certain probabilities.

Discrete random variables take on finitely many or countably infinitely many values. Their probability laws are often called **probability mass functions**. The following discrete random variables are frequently encountered.

Binomial

A **binomial** random variable X with parameters n and p has the probability law

$$\boxed{p_k = \Pr\{X = k\} = \binom{n}{k} p^k q^{n-k}} \tag{1.1}$$

$$\doteq b(k; n, p), \qquad k = 0, 1, 2, \ldots, n,$$

where $0 \leqslant p \leqslant 1$, $q = 1 - p$ and n is a positive integer ($\doteq$ means we are defining a new symbol). The **binomial coefficients** are

$$\binom{n}{k} = \frac{n!}{k!(n-k)!},$$

being the number of ways of choosing k items, without regard for order, from n distinguishable items.

When $n = 1$, so we have

$$\Pr\{X = 1\} = p = 1 - \Pr\{X = 0\},$$

the random variable is called **Bernoulli.**

Note the following.

Convention

Random variables are always designated by capital letters (e.g. X, Y) whereas symbols for the values they take on, as in Pr $\{X = k\}$, are always designated by lowercase letters.

The converse, however, is not true. Sometimes we use capital letters for non-random quantities.

Poisson

A **Poisson** random variable with parameter $\lambda > 0$ takes on non-negative integer values and has the probability law

$$\boxed{p_k = \Pr\{X = k\} = \frac{e^{-\lambda}\lambda^k}{k!}}, \qquad k = 0, 1, 2, \ldots. \tag{1.2}$$

For any random variable the total probability mass is unity. Hence if p_k is given by either (1.1) or (1.2),

$$\boxed{\sum_k p_k = 1}$$

where summation is over the possible values k as indicated.

For any random variable X, the **distribution function** is

$$F(x) = \Pr\{X \leqslant x\}, \quad -\infty < x < \infty.$$

Continuous random variables take on a continuum of values. Usually the probability law of a continuous random variable can be expressed through its **probability density function,** $f(x)$, which is the derivative of the distribution function. Thus

$$\begin{aligned} f(x) &= \frac{\mathrm{d}}{\mathrm{d}x} F(x) \\ &= \lim_{\Delta x \to 0} \frac{F(x+\Delta x) - F(x)}{\Delta x} \\ &= \lim_{\Delta x \to 0} \frac{\Pr\{X \leqslant x + \Delta x\} - \Pr\{X \leqslant x\}}{\Delta x} \\ &= \lim_{\Delta x \to 0} \frac{\Pr\{x < X \leqslant x + \Delta x\}}{\Delta x} \\ &= \lim_{\Delta x \to 0} \frac{\Pr\{X \in (x, x+\Delta x]\}}{\Delta x} \end{aligned} \tag{1.3}$$

The last two expressions in (1.3) often provide a convenient prescription for calculating probability density functions. Often the latter is abbreviated to p.d.f. but we will usually just say 'density'.

If the interval (x_1, x_2) is in the range of X then the probability that X takes values in this interval is obtained by integrating the probability density over (x_1, x_2).

$$\boxed{\Pr\{x_1 < X < x_2\} = \int_{x_1}^{x_2} f(x)\,\mathrm{d}x.}$$

The following continuous random variables are frequently encountered.

Normal (or Gaussian)

A random variable with density

$$\boxed{f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}}, \qquad -\infty < x < \infty, \tag{1.4}$$

$$\text{where} -\infty < \mu < \infty \quad \text{and} \quad 0 < \sigma^2 < \infty,$$

is called **normal**. The quantities μ and σ^2 are the mean and variance (elaborated upon below) and such a random variable is often designated

$N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$ the random variable is called a **standard normal** random variable, for which the usual symbol is Z.

Uniform

A random variable with constant density

$$\boxed{f(x) = \frac{1}{b-a}}, \qquad -\infty < a \leqslant x \leqslant b < \infty,$$

is said to be **uniformly distributed** on (a, b) and is denoted $U(a, b)$. If $a = 0, b = 1$ the density is unity on the unit interval,

$$f(x) = 1, \qquad 0 \leqslant x \leqslant 1$$

and the random variable is designated $U(0, 1)$.

Gamma

A random variable is said to have a **gamma density** (or gamma distribution) with parameters λ and ρ if

$$\boxed{f(x) = \frac{\lambda(\lambda x)^{\rho-1}e^{-\lambda x}}{\Gamma(\rho)}}, \qquad x \geqslant 0; \qquad \lambda, \rho > 0.$$

The quantity $\Gamma(\rho)$ is the **gamma function** defined as

$$\Gamma(\rho) = \int_0^\infty x^{\rho-1}e^{-x}\,\mathrm{d}x, \qquad \rho > 0.$$

When $\rho = 1$ the gamma density is that of an **exponentially distributed random variable**

$$\boxed{f(x) = \lambda e^{-\lambda x}}, \qquad x > 0.$$

For continuous random variables the density must integrate to unity:

$$\boxed{\int f(x)\,\mathrm{d}x = 1}$$

where the interval of integration is the whole range of values of X.

1.2 MEAN AND VARIANCE

Let X be a discrete random variable with

$$\Pr\{X = x_k\} = p_k, \qquad k = 1, 2, \ldots .$$

The **mean, average** or **expectation** of X is

$$\boxed{E(X) = \sum_k p_k x_k.}$$

For a binomial random variable $E(X) = np$ whereas a Poisson random variable has mean $E(X) = \lambda$.

For a continuous random variable with density $f(x)$,

$$\boxed{E(X) = \int x f(x)\, dx.}$$

If X is normal with density given by (1.4) then $E(X) = \mu$; a uniform (a, b) random variable has mean $E(X) = \frac{1}{2}(a + b)$; and a gamma variate has mean $E(X) = \rho/\lambda$.

The ***n*th moment** of X is the expected value of X^n:

$$E(X^n) = \begin{cases} \sum_k p_k x_k^n & \text{if } X \text{ is discrete,} \\ \int x^n f(x)\, dx & \text{if } X \text{ is continuous.} \end{cases}$$

If $n = 2$ we obtain the **second moment** $E(X^2)$. The **variance**, which measures the degree of dispersion of the probability mass of a random variable about its mean, is

$$\begin{aligned} \operatorname{Var}(X) &= E[(X - E(X))^2] \\ &= E(X^2) - E^2(X). \end{aligned}$$

The variances of the above-mentioned random variables are:

binomial, npq; Poisson, λ; normal, σ^2; uniform, $\frac{1}{12}(b - a)^2$; gamma, ρ/λ^2.

The square root of the variance is called the **standard deviation.**

1.3 CONDITIONAL PROBABILITY AND INDEPENDENCE

Let A and B be two random events. The **conditional probability** of A given B is, provided $\Pr\{B\} \neq 0$,

$$\boxed{\Pr\{A|B\} = \frac{\Pr\{AB\}}{\Pr\{B\}}}$$

where AB is the **intersection** of A and B, being the event that both A and B occur (sometimes written $A \cap B$). Thus only the occurrences of A which are simultaneous with those of B are taken into account. Similarly, if X, Y are random variables defined on the same sample space, taking on values $x_i, i = 1, 2, \ldots, y_j, j = 1, 2, \ldots$, then the conditional probability that $X = x_i$ given $Y = y_j$ is, if $\Pr\{Y = y_j\} \neq 0$,

$$\Pr\{X = x_i | Y = y_j\} = \frac{\Pr\{X = x_i, Y = y_j\}}{\Pr\{Y = y_j\}},$$

the comma between $X = x_i$ and $Y = y_j$ meaning 'and'.

The **conditional expectation** of X given $Y = y_j$ is

$$E(X | Y = y_j) = \sum_i x_i \Pr\{X = x_i | Y = y_j\}.$$

The **expected value** of XY is

$$E(XY) = \sum_{i,j} x_i y_j \Pr\{X = x_i, Y = y_j\},$$

and the **covariance** of X, Y is

$$\begin{aligned}\operatorname{Cov}(X, Y) &= E[(X - E(X))(Y - E(Y))] \\ &= E(XY) - E(X)E(Y).\end{aligned}$$

The covariance is a measure of the linear dependence of X on Y.

If X, Y are independent then the value of Y should have no effect on the probability that X takes on any of its values. Thus we define X, Y as **independent** if

$$\Pr\{X = x_i | Y = y_j\} = \Pr\{X = x_i\}, \qquad \text{all } i, j.$$

Equivalently X, Y are independent if

$$\Pr\{X = x_i, Y = y_j\} = \Pr\{X = x_i\} \Pr\{Y = y_j\},$$

with a similar formula for arbitrary independent events.

Hence for independent random variables

$$E(XY) = E(X)E(Y),$$

so their covariance is zero. Note, however, that $\operatorname{Cov}(X, Y) = 0$ does not always imply X, Y are independent. The covariance is often normalized by defining the **correlation coefficient**

$$\rho_{XY} = \frac{\operatorname{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

where σ_X, σ_Y are the standard deviations of X, Y. ρ_{XY} is bounded above and below by

$$\boxed{-1 \leqslant \rho_{XY} \leqslant 1}$$

Let $X_1, X_2, \ldots, X_n$ be **mutually independent** random variables. That is,

$$\begin{aligned}&\Pr\{X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n\}\\ &\quad = \Pr\{X_1 \in A_1\} \Pr\{X_2 \in A_2\} \ldots \Pr\{X_n \in A_n\},\end{aligned}$$

for all appropriate sets $A_1, \ldots, A_n$. Then

$$\boxed{\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \operatorname{Var}(X_i)}$$

so that variances add in the case of independent random variables. We also note the formula

$$\operatorname{Var}(aX + bY) = a^2 \operatorname{Var}(X) + b^2 \operatorname{Var}(Y),$$

which holds if X, Y are independent. If $X_1, X_2, \ldots, X_n$ are **independent identically distributed** (abbreviated to **i.i.d.**) random variables with $E(X_1) = \mu$, $\operatorname{Var}(X_1) = \sigma^2$, then

$$\mathrm{E}\left(\sum_{i=1}^{n} X_i\right) = \mu n; \qquad \operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = n\sigma^2.$$

If X is a random variable and $\{X_1, X_2, \ldots, X_n\}$ are i.i.d. with the distribution of X, then the collection $\{X_k\}$ is called a **random sample of size *n*** for X. Random samples play a key role in computer simulation (Chapter 5) and of course are fundamental in statistics.

1.4 LAW OF TOTAL PROBABILITY

Let Ω be a sample space for a random experiment and let $\{A_i, i = 1, 2, \ldots\}$ be a collection of nonempty subsets of Ω such that

(i) $A_i A_j = \varnothing, \qquad i \neq j;$
(ii) $\bigcup_i A_i = \Omega.$

(Here $\varnothing$ is the null set, the impossible event, being the complement of Ω.) Condition (i) says that the A_i represent **mutually exclusive** events. Condition (ii) states that when an experiment is performed, at least one of the A_i must be observed. Under these conditions the sets or events $\{A_i, i = 1, 2, \ldots\}$ are said to form a **partition** or **decomposition** of the sample space.

The **law or theorem of total probability** states that for any event (set) B,

$$\boxed{\Pr\{B\} = \sum_i \Pr\{B|A_i\} \Pr\{A_i\}}$$

A similar relation holds for expectations. By definition the expectation of X conditioned on the event A_i is

$$E(X|A_i) = \sum_k x_k \Pr\{X = x_k|A_i\},$$

where $\{x_k\}$ is the set of possible values of X. Thus

$$\begin{aligned} E(X) &= \sum_k x_k \Pr\{X = x_k\} \\ &= \sum_k x_k \sum_i \Pr\{X = x_k|A_i\} \Pr\{A_i\} \\ &= \sum_i \Pr\{A_i\} \sum_k x_k \Pr\{X = x_k|A_i\}. \end{aligned}$$

Thus

$$\boxed{E(X) = \sum_i E(X|A_i) \Pr\{A_i\}}$$

which we call the **law of total probability applied to expectations.**

We note also the fundamental relation for any two events A, B in the same sample space:

$$\boxed{\Pr\{A \cup B\} = \Pr\{A\} + \Pr\{B\} - \Pr\{AB\}}$$

where $A \cup B$ is the **union** of A and B, consisting of those points which are in A or in B or in both A and B.

1.5 CHANGE OF VARIABLES

Let X be a continuous random variable with distribution function F_X and density f_X. Let

$$y = g(x)$$

be a strictly increasing function of x (see Fig. 1.1) with inverse function

$$x = h(y).$$

Then

$$Y = g(X)$$

is a random variable which we let have distribution function F_Y and density f_Y.

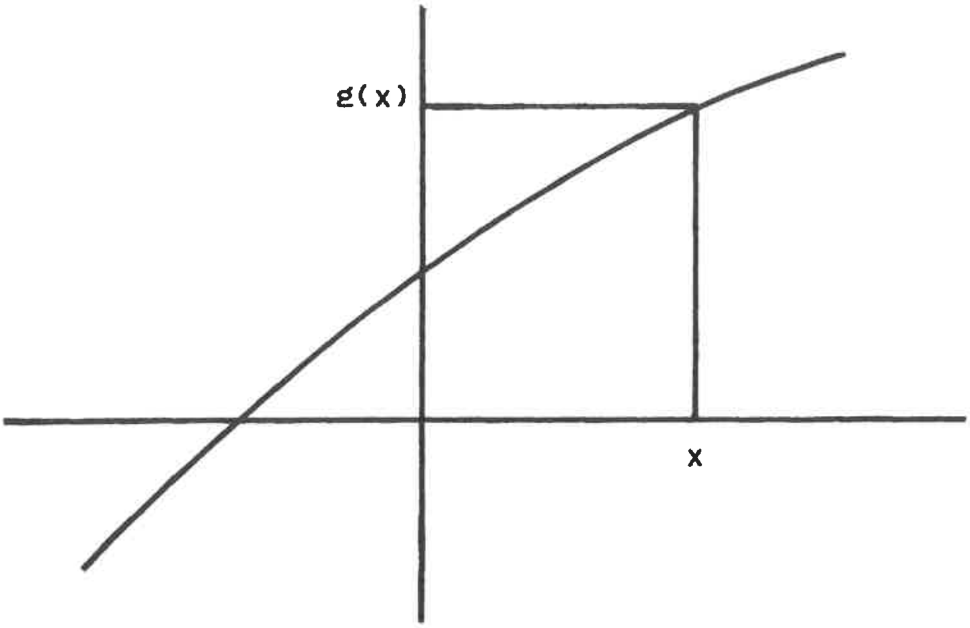

Figure 1.1 $g(x)$ is a strictly increasing function of x.

It is easy to see that $X \leqslant x$ implies $Y \leqslant g(x)$. Hence we arrive at

$$\Pr\{X \leqslant x\} = \Pr\{Y \leqslant g(x)\}$$

By the definition of distribution functions this can be written

$$F_X(x) = F_Y(g(x)). \tag{1.5}$$

Therefore

$$F_Y(y) = F_X(h(y)).$$

On differentiating with respect to y we obtain, assuming that h is differentiable,

$$\frac{\mathrm{d}F_Y}{\mathrm{d}y} = \left.\frac{\mathrm{d}F_X(x)}{\mathrm{d}x}\right|_{h(y)} \frac{\mathrm{d}h}{\mathrm{d}y}$$

or in terms of densities

$$f_Y(y) = f_X(h(y))\frac{\mathrm{d}h}{\mathrm{d}y}. \tag{1.6}$$

If y is a strictly decreasing function of x we obtain

$$\Pr\{X \leqslant x\} = \Pr\{Y \geqslant g(x)\}.$$

Working through the steps between (1.5) and (1.6) in this case gives

$$f_Y(y) = f_X(h(y))\left(-\frac{\mathrm{d}h}{\mathrm{d}y}\right). \tag{1.7}$$

Both formulas (1.6) and (1.7) are covered by the single formula

$$\boxed{f_Y(y) = f_X(h(y))\left|\frac{dh}{dy}\right|}$$

where | | denotes **absolute value.** Cases where g is neither strictly increasing nor strictly decreasing require special consideration.

1.6 TWO-DIMENSIONAL RANDOM VARIABLES

Let X, Y be random variables defined on the same sample space. Then their **joint distribution function** is

$$F_{XY}(x, y) = \Pr\{X \leqslant x, Y \leqslant y\}.$$

The mixed partial derivative of F_{XY}, if it exists, is the **joint density** of X and Y:

$$f_{XY}(x, y) = \frac{\partial^2 F_{XY}}{\partial x \partial y}.$$

As a rough guide we have, for small enough $\Delta x, \Delta y$,

$$f_{XY}(x, y)\Delta x \Delta y \simeq \Pr\{X \in (x, x + \Delta x], \mathrm{Y} \in (y, y + \Delta y]\}.$$

If X, Y are independent then their joint distribution function and joint density function factor into those of the individual random variables:

$$F_{XY}(x, y) = F_X(x)F_Y(y),$$
$$f_{XY}(x, y) = f_X(x)f_Y(y).$$

In particular, if X, Y are **independent standard normal random variables,**

$$f_{XY}(x, y) = \left(\frac{1}{\sqrt{2\pi}}\exp\left\{\frac{-x^2}{2}\right\}\right)\left(\frac{1}{\sqrt{2\pi}}\exp\left\{\frac{-y^2}{2}\right\}\right), \qquad -\infty < x, y < \infty,$$

which can be written

$$\boxed{f_{XY}(x, y) = \frac{1}{2\pi}\exp\{-\tfrac{1}{2}(x^2 + y^2)\}} \tag{1.8}$$

In fact if the joint density X, Y is as given by (1.8) we may conclude that X, Y are independent standard normal random variables.

Change of variables

Let U, V be random variables with joint density $f_{UV}(u, v)$. Suppose that the one–one mappings means that 5% of the time, values of χ^2 greater than the critical value occur even when H_0 is true. That is, there is a 5% chance that we will (incorrectly) reject H_0 when it is true.

In applying the above χ^2 goodness of fit test, the number of degrees of freedom is given by the number n, of 'cells', minus the number of linear relations between the N_i. (There is at least one, $\sum N_i = N$.) The number of degrees of freedom is reduced further by one for each estimated parameter needed to describe the distribution under H_0.

It is recommended that the expected numbers of observations in each category should not be less than 5, but this requirement can often be relaxed. A table of critical values of χ^2 is given in the Appendix, p. 219.

For a detailed account of hypothesis testing and introductory statistics generally, see for example Walpole and Myers (1985), Hogg and Craig (1978) and Mendenhall, Scheaffer and Wackerly (1981). For full accounts of basic probability theory see also Chung (1979) and Feller (1968). Two recent books on applications of probability at an undergraduate level are those of Ross (1985) and Taylor and Karlin (1984).

1.8 NOTATION

Little o

A quantity which depends on Δx but vanishes more quickly than Δx as $\Delta x \to 0$ is said to be 'little o of Δx', written $o(\Delta x)$. Thus for example $(\Delta x)^2$ is $o(\Delta x)$ because $(\Delta x)^2$ vanishes more quickly than Δx. In general, if

$$\lim_{\Delta x \to 0} \frac{g(\Delta x)}{\Delta x} = 0,$$

we write

$$g(\Delta x) = o(\Delta x).$$

The little o notation is very useful to abbreviate expressions in which terms will not contribute after a limiting operation is taken. To illustrate, consider the Taylor expansion of $e^{\Delta x}$:

$$\begin{aligned} e^{\Delta x} &= 1 + \Delta x + \frac{(\Delta x)^2}{2!} + \frac{(\Delta x)^3}{3!} + \cdots \\ &= 1 + \Delta x + o(\Delta x). \end{aligned}$$

We then have

$$\left.\frac{\mathrm{d}}{\mathrm{d}x} e^x \right|_{x=0} = \lim_{\Delta x \to 0} \frac{e^{\Delta x} - 1}{\Delta x}$$

$$= \lim_{\Delta x \to 0} \frac{1 + \Delta x + o(\Delta x) - 1}{\Delta x}$$

$$= \lim_{\Delta x \to 0} \frac{\Delta x}{\Delta x} + \frac{o(\Delta x)}{\Delta x}$$

$$= 1.$$

Equal by definition

As seen already, when we write, for example,

$$q \doteq (1 - p)$$

we are defining the symbol q to be equal to $1 - p$. This is not to be confused with approximately equal to, which is indicated by $\simeq$.

Unit step function

The **unit** (or **Heaviside**) step function located at x_0 is

$$H(x - x_0) = \begin{cases} 0, & x < x_0, \\ 1, & x \geqslant x_0. \end{cases}$$

Thus $H(x - x_0)$ has a jump of $+1$ at x_0 and it is **right-continuous.**

i.i.d.

As seen already, the letters i.i.d. stand for independent and identically distributed.

Probability

Usually the probability of an event A is written

$$\Pr\{A\}$$

but occasionally we just write

$$P\{A\}.$$

REFERENCES

Blake, I.F. (1979). *An Introduction to Applied Probability.* Wiley, New York.

Chung, K.L. (1979). *Elementary Probability Theory.* Springer-Verlag, New York.

Feller, W. (1968). *An Introduction to Probability Theory and its Applications.* Wiley, New York.

Hogg, R.V. and Craig, A.T. (1978). *Introduction to Mathematical Statistics.* Macmillan, New York.

Mendenhall, W., Scheaffer, R.L. and Wackerly, D.D. (1981). *Mathematical Statistics with Applications.* Duxbury, Boston.