1

Random variables

1.1 Random variables

Statistics is about extracting information from data that contain an inherently unpredictable component. *Random variables* are the mathematical construct used to build models of such variability. A random variable takes a different value, at random, each time it is observed. We cannot say, in advance, exactly what value will be taken, but we can make probability statements about the values likely to occur. That is, we can characterise the *distribution* of values taken by a random variable. This chapter briefly reviews the technical constructs used for working with random variables, as well as a number of generally useful related results. See De Groot and Schervish (2002) or Grimmett and Stirzaker (2001) for fuller introductions.

1.2 Cumulative distribution functions

The cumulative distribution function (c.d.f.) of a random variable (r.v.), X, is the function $F(x)$ such that

$$F(x) = \Pr(X \leq x).$$

That is, $F(x)$ gives the probability that the value of X will be less than or equal to x. Obviously, $F(-\infty) = 0$, $F(\infty) = 1$ and $F(x)$ is monotonic. A useful consequence of this definition is that if F is continuous then $F(X)$ has a uniform distribution on $[0, 1]$: it takes any value between 0 and 1 with equal probability. This is because

$$\Pr(X \leq x) = \Pr\{F(X) \leq F(x)\} = F(x) \Rightarrow \Pr\{F(X) \leq u\} = u$$

(if F is continuous), the latter being the c.d.f. of a uniform r.v. on $[0, 1]$.

Define the inverse of the c.d.f. as $F^-(u) = \min(x|F(x) \geq u)$, which is just the usual inverse function of F if F is continuous. F^- is often called the *quantile function* of X. If U has a uniform distribution on $[0, 1]$, then

$F^-(U)$ is distributed as X with c.d.f. F. Given some way of generating uniform random deviates, this provides a method for generating random variables from any distribution with a computable F^-.

Let p be a number between 0 and 1. The p *quantile* of X is the value that X will be less than or equal to, with probability p. That is, $F^-(p)$. Quantiles have many uses. One is to check whether data, $x_1, x_2, \ldots, x_n$, could plausibly be observations of a random variable with c.d.f. F. The x_i are sorted into order, so that they can be treated as 'observed quantiles'. They are then plotted against the theoretical quantiles $F^-\{(i-0.5)/n\}$ $(i = 1, \ldots, n)$ to produce a *quantile-quantile plot* (QQ-plot). An approximately straight-line QQ-plot should result, if the observations are from a distribution with c.d.f. F.

1.3 Probability (density) functions

For many statistical methods a function that tells us about the probability of a random value taking a particular value is more useful than the c.d.f. To discuss such functions requires some distinction to be made between random variables taking a discrete set of values (e.g. the non-negative integers) and those taking values from intervals on the real line.

For a discrete random variable, X, the *probability function* (or probability *mass* function), $f(x)$, is the function such that

$$f(x) = \Pr(X = x).$$

Clearly $0 \leq f(x) \leq 1$, and since X must take some value, $\sum_i f(x_i) = 1$, where the summation is over all possible values of x (denoted x_i).

Because a continuous random variable, X, can take an infinite number of possible values, the probability of taking any particular value is usually zero, so that a probability function would not be very useful. Instead the *probability density function*, $f(x)$, gives the probability per unit interval of X being near x. That is, $\Pr(x - \Delta/2 < X < x + \Delta/2) \simeq f(x)\Delta$. More formally, for any constants $a \leq b$,

$$\Pr(a \leq X \leq b) = \int_a^b f(x)dx.$$

Clearly this only works if $f(x) \geq 0$ and $\int_{-\infty}^{\infty} f(x)dx = 1$. Note that $\int_{-\infty}^{b} f(x)dx = F(b)$, so $F'(x) = f(x)$ when F' exists. Appendix A provides some examples of useful standard distributions and their probability (density) functions.

The following sections mostly consider continuous random variables, but except where noted, equivalent results also apply to discrete random variables upon replacement of integration by an appropriate summation. For conciseness the convention is adopted that p.d.f.s with different arguments usually denote different functions (e.g. $f(y)$ and $f(x)$ denote different p.d.f.s).

1.4 Random vectors

Little can usually be learned from single observations. Useful statistical analysis requires multiple observations and the ability to deal simultaneously with multiple random variables. A multivariate version of the p.d.f. is required. The two-dimensional case suffices to illustrate most of the required concepts, so consider random variables X and Y.

The *joint probability density function* of X and Y is the function $f(x, y)$ such that, if Ω is any region in the $x - y$ plane,

$$\Pr\{(X, Y) \in \Omega\} = \iint_{\Omega} f(x, y) dx dy. \tag{1.1}$$

So $f(x, y)$ is the probability *per unit area* of the $x - y$ plane, at x, y. If ω is a small region of area α, containing a point x, y, then $\Pr\{(X, Y) \in \omega\} \simeq f_{xy}(x, y)\alpha$. As with the univariate p.d.f. $f(x, y)$ is non-negative and integrates to one over $\mathbb{R}^2$.

Example Figure 1.1 illustrates the following joint p.d.f.

$$f(x, y) = \begin{cases} x + 3y^2/2 & 0 < x < 1 \ \& \ 0 < y < 1 \\ 0 & \text{otherwise.} \end{cases} \tag{1.2}$$

Figure 1.2 illustrates evaluation of two probabilities using this p.d.f.

1.4.1 Marginal distribution

Continuing with the X, Y case, the p.d.f. of X or Y, ignoring the other variable, can be obtained from $f(x, y)$. To find the *marginal* p.d.f. of X, we seek the probability density of X given that $-\infty < Y < \infty$. From the defining property of a p.d.f., it is unsurprising that this is

$$f(x) = \int_{-\infty}^{\infty} f(x, y) dy,$$

with a similar definition for $f(y)$.

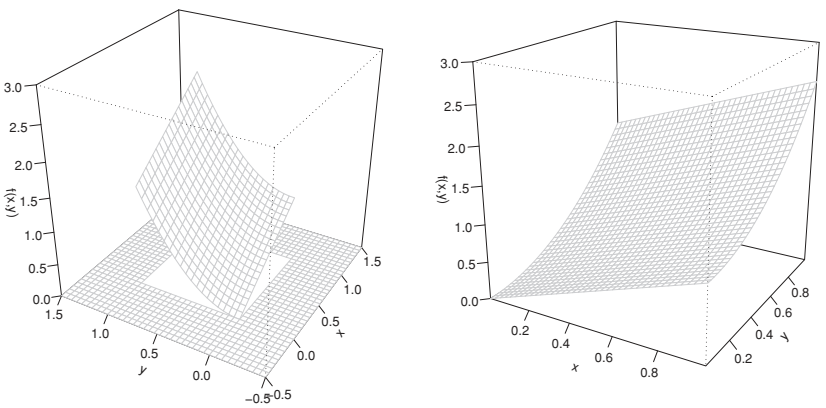

Figure 1.1 The example p.d.f (1.2). Left: over the region $[-0.5, 1.5] \times [-0.5, 1.5]$. Right: the nonzero part of the p.d.f.

1.4.2 Conditional distribution

Suppose that we know that Y takes some particular value y_0. What does this tell us about the distribution of X? Because X and Y have joint density $f(x, y)$, we would expect the density of x, given $Y = y_0$, to be proportional to $f(x, y_0)$. That is, we expect

$$f(x|Y = y_0) = kf(x, y_0),$$

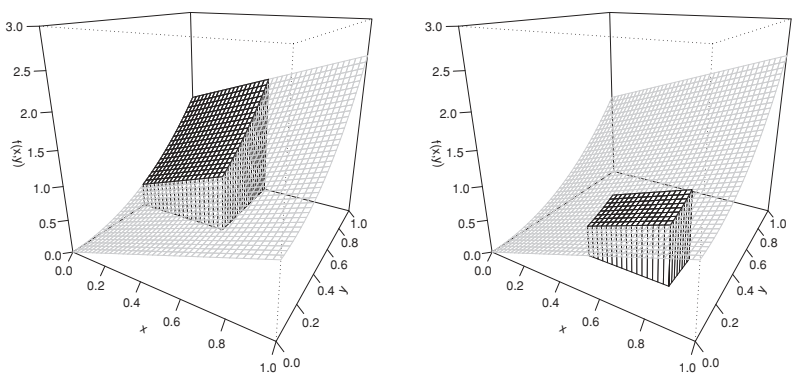

Figure 1.2 Evaluating probabilities from the joint p.d.f. (1.2), shown in grey. Left: in black is shown the volume evaluated to find $\Pr[X < .5, Y > .5]$. Right: $\Pr[.4 < X < .8, .2 < Y < .4]$.

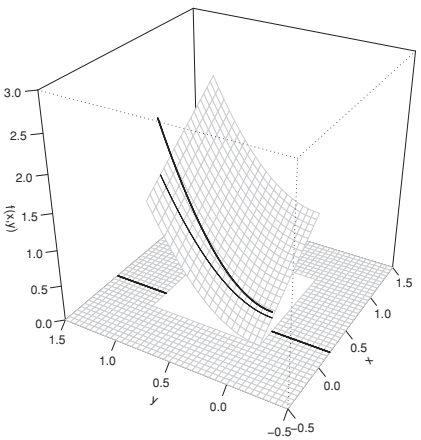

Figure 1.3 The conditional density $f(y|.2)$. The joint density $f(x, y)$ is shown as a grey surface. The thin black curve shows $f(.2, y)$. The thick black curve shows $f(y|.2) = f(.2, y)/f_x(.2)$.

where k is a constant. Now if $f(x|y)$ is a probability density function, then it must integrate to 1. So,

$$k\int_{-\infty}^{\infty} f(x, y_0)dx = 1 \quad \Rightarrow \quad kf(y_0) = 1 \Rightarrow k = \frac{1}{f(y_0)},$$

where $f(y_0)$ denotes the marginal density of y at y_0. Hence we have:

Definition If X and Y have joint density $f(x, y)$ then the *conditional density* of X, given $Y = y_0$, is

$$f(x|Y = y_0) = \frac{f(x, y_0)}{f(y_0)}, \tag{1.3}$$

assuming $f(y_0) > 0$.

Notice that this is a p.d.f. for random variable X: y_0 is now fixed. To simplify notation we can also write $f(x|y_0)$ in place of $f(x|Y = y_0)$, when the meaning is clear. Of course, symmetric definitions apply to the conditional distribution of Y given X: $f(y|x_0) = f(x_0, y)/f(x_0)$. Figure 1.3 illustrates the relationship between joint and conditional p.d.f.s.

Manipulations involving the replacement of joint distributions with conditional distributions, using $f(x, y) = f(x|y)f(y)$, are common in

statistics, but not everything about generalising beyond two dimensions is completely obvious, so the following three examples may help.

1. $f(x, z|y) = f(x|z, y)f(z|y)$.
2. $f(x, z, y) = f(x|z, y)f(z|y)f(y)$.
3. $f(x, z, y) = f(x|z, y)f(z, y)$.

1.4.3 Bayes theorem

From the previous section it is clear that

$$f(x, y) = f(x|y)f(y) = f(y|x)f(x).$$

Rearranging the last two terms gives

$$f(x|y) = \frac{f(y|x)f(x)}{f(y)}.$$

This important result, *Bayes theorem*, leads to a whole school of statistical modelling, as we see in chapters 2 and 6.

1.4.4 Independence and conditional independence

If random variables X and Y are such that $f(x|y)$ does not depend on the value of y, then x is statistically *independent* of y. This has the consequence that

$$\begin{aligned} f(x) = \int_{-\infty}^{\infty} f(x, y)dy &= \int_{-\infty}^{\infty} f(x|y)f(y)dy \\ &= f(x|y)\int_{-\infty}^{\infty} f(y)dy = f(x|y), \end{aligned}$$

which in turn implies that $f(x, y) = f(x|y)f(y) = f(x)f(y)$. Clearly the reverse implication also holds, since $f(x, y) = f(x)f(y)$ leads to $f(x|y) = f(x, y)/f(y) = f(x)f(y)/f(y) = f(x)$. In general then:

Random variables X and Y are independent if and only if their joint p.(d.)f. is given by the product of their marginal p.(d.)f.s: that is, $f(x, y) = f(x)f(y)$.

Modelling the elements of a random vector as independent usually simplifies statistical inference. Assuming independent *identically distributed* (i.i.d.) elements is even simpler, but much less widely applicable.

In many applications, a set of observations cannot be modelled as independent, but can be modelled as *conditionally independent*. Much of modern statistical research is devoted to developing useful models that exploit various sorts of conditional independence in order to model dependent data in computationally feasible ways.

Consider a sequence of random variables $X_1, X_2, \ldots X_n$, and let $\mathbf{X}_{-i} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)^{\mathrm{T}}$. A simple form of conditional independence is the first order Markov property,

$$f(x_i|\mathbf{x}_{-i}) = f(x_i|x_{i-1}).$$

That is, X_{i-1} completely determines the distribution of X_i, so that *given* X_{i-1}, X_i is independent of the rest of the sequence. It follows that

$$\begin{aligned} f(\mathbf{x}) = f(x_n|\mathbf{x}_{-n})f(\mathbf{x}_{-n}) &= f(x_n|x_{n-1})f(\mathbf{x}_{-n}) \\ &= \ldots = \prod_{i=2}^{n} f(x_i|x_{i-1})f(x_1), \end{aligned}$$

which can often be exploited to yield considerable computational savings.

1.5 Mean and variance

Although it is important to know how to characterise the distribution of a random variable completely, for many purposes its first- and second-order properties suffice. In particular the *mean* or *expected value* of a random variable, X, with p.d.f. $f(x)$, is defined as

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx.$$

Since the integral is weighting each possible value of x by its relative frequency of occurrence, we can interpret $E(X)$ as being the average of an infinite sequence of observations of X.

The definition of expectation applies to any function g of X:

$$E\{g(X)\} = \int_{-\infty}^{\infty} g(x)f(x)dx.$$

Defining $\mu = E(X)$, then a particularly useful g is $(X - \mu)^2$, measuring the squared difference between X and its average value, which is used to define the *variance* of X:

$$\mathrm{var}(X) = E\{(X - \mu)^2\}.$$

The variance of X measures how spread out the distribution of X is. Although computationally convenient, its interpretability is hampered by having units that are the square of the units of X. The *standard deviation* is the square root of the variance, and hence is on the same scale as X.

1.5.1 Mean and variance of linear transformations

From the definition of expectation it follows immediately that if a and b are finite real constants $E(a + bX) = a + bE(X)$. The variance of $a + bX$ requires slightly more work:

$$\begin{aligned}\text{var}(a + bX) &= E\{(a + bX - a - b\mu)^2\} \\ &= E\{b^2(X - \mu)^2\} = b^2 E\{(X - \mu)^2\} = b^2\text{var}(X).\end{aligned}$$

If X and Y are random variables then $E(X + Y) = E(X) + E(Y)$. To see this suppose that they have joint density $f(x, y)$; then,

$$\begin{aligned}E(X + Y) &= \int (x + y) f(x, y) dxdy \\ &= \int x f(x, y) dxdy + \int y f(x, y) dxdy = E(X) + E(Y).\end{aligned}$$

This result assumes nothing about the distribution of X and Y. If we now add the assumption that X and Y are independent then we find that $E(XY) = E(X)E(Y)$ as follows:

$$\begin{aligned}E(XY) &= \int xy f(x, y) dxdy \\ &= \int x f(x) y f(y) dxdy \ \text{ (by independence)} \\ &= \int x f(x) dx \int y f(y) dy = E(X)E(Y).\end{aligned}$$

Note that the reverse implication only holds if the joint distribution of X and Y is Gaussian.

Variances do not add as nicely as means (unless X and Y are independent), and we need the notion of *covariance*:

$$\text{cov}(X, Y) = E\{(X - \mu_x)(Y - \mu_y)\} = E(XY) - E(X)E(Y),$$

where $\mu_x = E(X)$ and $\mu_y = E(Y)$. Clearly $\text{var}(X) \equiv \text{cov}(X, X)$, and if X and Y are independent $\text{cov}(X, Y) = 0$ (since then $E(XY) = E(X)E(Y)$).

Now let $\mathbf{A}$ and $\mathbf{b}$ be, respectively, a matrix and a vector of fixed finite coefficients, with the same number of rows, and let $\mathbf{X}$ be a random vector. $E(\mathbf{X}) = \boldsymbol{\mu}_x = \{E(X_1), E(X_2), \ldots, E(X_n)\}^{\mathrm{T}}$ and it is immediate that $E(\mathbf{AX} + \mathbf{b}) = \mathbf{A}E(\mathbf{X}) + \mathbf{b}$. A useful summary of the second-order properties of $\mathbf{X}$ requires both variances and covariances of its elements. These can be written in the (symmetric) variance-covariance matrix $\boldsymbol{\Sigma}$, where $\Sigma_{ij} = \mathrm{cov}(X_i, X_j)$, which means that

$$\boldsymbol{\Sigma} = E\{(\mathbf{X} - \boldsymbol{\mu}_x)(\mathbf{X} - \boldsymbol{\mu}_x)^{\mathrm{T}}\}. \tag{1.4}$$

A very useful result is that

$$\boldsymbol{\Sigma}_{AX+b} = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^{\mathrm{T}}, \tag{1.5}$$

which is easily proven:

$$\begin{aligned}\boldsymbol{\Sigma}_{AX+b} &= E\{(\mathbf{AX} + \mathbf{b} - \mathbf{A}\boldsymbol{\mu}_x - \mathbf{b})(\mathbf{AX} + \mathbf{b} - \mathbf{A}\boldsymbol{\mu}_x - \mathbf{b})^{\mathrm{T}}\} \\ &= E\{(\mathbf{AX} - \mathbf{A}\boldsymbol{\mu}_x)(\mathbf{AX} - \mathbf{A}\boldsymbol{\mu}_x)^{\mathrm{T}})\} \\ &= \mathbf{A}E\{(\mathbf{X} - \boldsymbol{\mu}_x)(\mathbf{X} - \boldsymbol{\mu}_x)^{\mathrm{T}}\}\mathbf{A}^{\mathrm{T}} = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^{\mathrm{T}}.\end{aligned}$$

So if $\mathbf{a}$ is a vector of fixed real coefficients then $\mathrm{var}(\mathbf{a}^{\mathrm{T}}\mathbf{X}) = \mathbf{a}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{a} \geq 0$: a covariance matrix is positive semi-definite.

1.6 The multivariate normal distribution

The normal or *Gaussian* distribution (see Section A.1.1) has a central place in statistics, largely as a result of the central limit theorem covered in Section 1.9. Its multivariate version is particularly useful.

Definition Consider a set of n i.i.d. standard normal random variables: $Z_i \underset{\text{i.i.d}}{\sim} N(0, 1)$. The covariance matrix for $\mathbf{Z}$ is $\mathbf{I}_n$ and $E(\mathbf{Z}) = \mathbf{0}$. Let $\mathbf{B}$ be an $m \times n$ matrix of fixed finite real coefficients and $\boldsymbol{\mu}$ be an m- vector of fixed finite real coefficients. The m-vector $\mathbf{X} = \mathbf{BZ} + \boldsymbol{\mu}$ is said to have a *multivariate normal distribution*. $E(\mathbf{X}) = \boldsymbol{\mu}$ and the covariance matrix of $\mathbf{X}$ is just $\boldsymbol{\Sigma} = \mathbf{B}\mathbf{B}^{\mathrm{T}}$. The short way of writing $\mathbf{X}$'s distribution is

$$\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

In Section 1.7, basic transformation theory establishes that the p.d.f. for this distribution is

$$f_{\mathbf{x}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^m |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \quad \text{for} \quad \mathbf{x} \in \mathbb{R}^m, \tag{1.6}$$

assuming $\boldsymbol{\Sigma}$ has full rank (if $m = 1$ the definition gives the usual univariate normal p.d.f.). Actually there exists a more general definition in which $\boldsymbol{\Sigma}$ is merely positive semi-definite, and hence potentially singular: this involves a *pseudoinverse* of $\boldsymbol{\Sigma}$.

An interesting property of the multivariate normal distribution is that if X and Y have a multivariate normal distribution and zero covariance, then they must be independent. This implication only holds for the normal (independence implies zero covariance for any distribution).

1.6.1 A multivariate t distribution

If we replace the random variables $Z_i \underset{\text{i.i.d}}{\sim} N(0,1)$ with random variables $T_i \underset{\text{i.i.d}}{\sim} t_k$ (see Section A.1.3) in the definition of a multivariate normal, we obtain a vector with a multivariate $t_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. This can be useful in stochastic simulation, when we need a multivariate distribution with heavier tails than the multivariate normal. Note that the resulting univariate marginal distributions are not t distributed. Multivariate t densities with t distributed marginals are more complicated to characterise.

1.6.2 Linear transformations of normal random vectors

From the definition of multivariate normality, it immediately follows that if $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{A}$ is a matrix of finite real constants (of suitable dimensions), then

$$\mathbf{AX} \sim N(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^{\mathrm{T}}). \tag{1.7}$$

This is because $\mathbf{X} = \mathbf{BZ} + \boldsymbol{\mu}$, so $\mathbf{AX} = \mathbf{ABZ} + \mathbf{A}\boldsymbol{\mu}$, and hence $\mathbf{AX}$ is exactly the sort of linear transformation of standard normal r.v.s that defines a multivariate normal random vector. Furthermore it is clear that $E(\mathbf{AX}) = \mathbf{A}\boldsymbol{\mu}$ and the covariance matrix of $\mathbf{AX}$ is $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^{\mathrm{T}}$.

A special case is that if $\mathbf{a}$ is a vector of finite real constants, then

$$\mathbf{a}^{\mathrm{T}}\mathbf{X} \sim N(\mathbf{a}^{\mathrm{T}}\boldsymbol{\mu}, \mathbf{a}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{a}).$$

For the case in which $\mathbf{a}$ is a vector of zeros, except for a_j, which is 1, (1.7) implies that

$$X_j \sim N(\mu_j, \Sigma_{jj}) \tag{1.8}$$

(usually we would write σ_j^2 for Σ_{jj}). In words:

If $\mathbf{X}$ has a multivariate normal distribution, then the marginal distribution of any X_j is univariate normal.