BIVARIATE DISTRIBUTIONS

Let x be a variable that assumes the values $\{x_1, x_2, \ldots, x_n\}$. Then, a function that expresses the relative frequency of these values is called a univariate frequency function. It must be true that

$$f(x_i) \geq 0 \quad \text{for all} \quad i \qquad \text{and} \qquad \sum_i f(x_i) = 1.$$

The following table provides a trivial example:

x	$f(x)$
-1	0.25
1	0.75

Let x and y be variables that assume values in the the sets $\{x_1, x_2, \ldots, x_n\}$ and $\{y_1, y_2, \ldots, y_m\}$, respectively. Then the function $f(x_i, y_j)$, which gives the relative frequencies of the occurrence of the pairs (x_i, y_j) is called a bivariate frequency function. It must be true that

$$f(x_i, y_j) \geq 0 \quad \text{for all} \quad i \qquad \text{and} \qquad \sum_i \sum_j f(x_i, y_j) = 1.$$

An example of a bivariate frequency table is as follows:

		y	
x	-1	0	1
-1	0.04	0.01	0.20
1	0.12	0.03	0.60

The values of $f(x_i, y_j)$ are within the body of the table.

The marginal frequency function of x gives the relative frequencies of the values of x_i regardless of the values of y_j with which they are associated; and it is defined by

$$f(x_i) = \sum_j f(x_i, y_j); \quad i = 1, \ldots, n.$$

It follows that

$$f(x_i) \geq 0, \quad \text{and} \quad \sum_i f(x_i) = \sum_i \sum_j f(x_i, y_j) = 1,$$

The marginal frequency function $f(y_j)$ is defined analogously.

The bivariate frequency table above provides examples of the two marginal frequency functions:

$$f(x = -1) = 0.04 + 0.01 + 0.20 = 0.25,$$

$$f(x = 1) = 0.12 + 0.03 + 0.60 = 0.75.$$

and

$$f(y=-1) = 0.04 + 0.12 = 0.16,$$
$$f(y=0) = 0.01 + 0.03 = 0.04,$$
$$f(y=1) = 0.20 + 0.60 = 0.80.$$

The conditional frequency function of x given $y = y_j$ gives the relative frequency of the values of x_i in the subset of $f(x, y)$ for which $y = y_j$; and it is given by

$$f(x_i|y_j) = \frac{f(x_i, y_j)}{f(y_j)}.$$

Observe that

$$\sum_i f(x_i|y_j) = \frac{\sum_i f(x_i, y_j)}{f(y_j)} = \frac{f(y_j)}{f(y_j)} = 1.$$

An example based on the bivariate table is as follows:

$f(x|y)$

x	$f(x\|y=-1)$	$f(x\|y=0)$	$f(x\|y=1)$
-1	$0.25 = (0.04/0.16)$	$0.25 = (0.01/0.04)$	$0.25 = (0.20/0.80)$
1	$0.75 = (0.12/0.16)$	$0.75 = (0.03/0.04)$	$0.75 = (0.60/0.80)$

We may say that x is independent of y if and only if the conditional distribution of x is the same for all values of y, as it is in this table.

The conditional frequency functions of x are the same for all values of y if and only if they are all equal to the marginal frequency function of x.

Proof. Suppose that $f^*(x_i) = f(x|y_1) = \cdots = f(x|y_m)$. Then

$$f(x_i) = \sum_j f(x_i|y_j)f(y_j) = f^*(x_i)\sum_j f(y_j) = f^*(x_i),$$

which is to say that $f^*(x_i) = f(x_i)$, Conversely, if the conditionals are all equal to the marginal, then they must be equal to each other.

Also observe that, if $f(x_i|y_j) = f(x_i)$ for all j and $f(y_j|x_i) = f(y_j)$ for all i, then, equivalently,

$$f(x_i, y_j) = f(x_i|y_j)f(y_j) = f(y_j|x_i)f(x_i) = f(x_i)f(y_j).$$

The condition that $f(x_i, y_j) = f(x_i)f(y_j)$ constitutes an equivalent definition of the independence of x and y.

We have been concerned, so far, with frequency functions. These provide the prototypes for bivariate probability mass functions and for bivariate probability density functions. The extension to probability mass functions is immediate. For the case of the density functions, we consider a two-dimensional space $\mathcal{R}^2$ which is defined as the set of all ordered pairs $(x, y); -\infty < x, y < \infty$, which correspond to the co-ordinates of the points in a plane of infinite extent.

We suppose that there is a probability measure defined over $\mathcal{R}^2$ such that, for any $\mathcal{A} \subset \mathcal{R}^2$, $P(\mathcal{A})$ is the probability that (x, y) falls in $\mathcal{A}$. Thus, for example, if $\mathcal{A} = \{a < x \leq b, a < x \leq b\}$, which is a rectangle in the plane, then

$$P(\mathcal{A}) = \int_{y=c}^{d} \left\{ \int_{x=a}^{b} f(x, y)dx \right\} dy.$$

This is a double integral, which is performed in respect of the two variables in succession and in either order. Usually, the braces are omitted, which is allowable if care is taken to ensure the correct correspondence between the integral signs and the differentials.

Example. Let (x, y) be a random vector with a p.d.f of

$$f(x, y) = \frac{1}{8}(6 - x - y); \quad 0 \leq x \leq 2; \quad 2 \leq y \leq 4.$$

It needs to be confirmed that this does integrate to unity over the specified range of (x, y). There is

$$\begin{aligned} \frac{1}{8} \int_{x=0}^{2} \int_{y=2}^{4} (6 - x - y)dydx &= \frac{1}{8} \int_{x=0}^{2} \left[6y - xy - \frac{y^2}{2} \right]_2^4 dx \\ &= \frac{1}{8} \int_{x=0}^{2} (6 - 2x)\, dx = \frac{1}{8} \left[6x - x^2 \right]_0^2 = \frac{8}{8} = 1. \end{aligned}$$

Moments of a bivariate distribution. Let (x, y) have the p.d.f. $f(x, y)$. Then, the expected value of x is defined by

$$E(x) = \int_x \int_y x f(x, y)dydx = \int_x x f(x)dx, \quad \text{if } x \text{ is continuous, and by}$$

$$E(x) = \sum_x \sum_y x f(x, y) = \sum_x x f(x), \quad \text{if } x \text{ is discrete.}$$

Joint moments of x and y can also be defined. For example, there is

$$\int_x \int_y (x - a)^r (y - b)^s f(x, y)dydx,$$

where r, s are integers and a, b are fixed data. The most important joint moment for present purposes is the covariance of x and y, defined by

$$C(x, y) = \int_x \int_y \{x - E(x)\}\{y - E(y)\} f(x, y)dydx.$$

If x, y are statistically independent, such that $f(x, y) = f(x)f(y)$, then their joint moments can be expressed as the products of their separate moments.

Thus, for example, if x, y are independent then

$$E\{[x - E(x)]^2[y - E(y)]^2\} = \int_x \int_y [x - E(x)]^2[y - E(y)]^2 f(x)f(y)dydx$$
$$= \int_x [x - E(x)]^2 f(x)dx \int_y [y - E(y)]^2 f(y)dy = V(x)V(y).$$

The case of the covariance of x, y, when these are independent, is of prime importance:

$$C(x, y) = E\{[x - E(x)][y - E(y)]\} = \int_x [x - E(x)]f(x)dx \int_y [y - E(y)]f(y)dy$$
$$= \{[E(x) - E(x)][E(y) - E(y)]\} = 0.$$

These relationships are best expressed using the notation of the expectations operator. Thus

$$C(x, y) = E\{[x - E(x)][y - E(y)]\} = E[xy - E(x)y - xE(y) + E(x)E(y)]$$
$$= E(xy) - E(x)E(y) - E(x)E(y) + E(x)E(y)$$
$$= E(xy) - E(x)E(y).$$

Since $E(xy) = E(x)E(y)$ if x, y are independent, it follows, in that case, that $C(x, y) = 0$. Observe also that $C(x, x) = E\{[x - E(x)]^2\} = V(x)$.

Now consider the variance of the sum $x + y$. This is

$$V(x + y) = E\left\{[(x + y) - E(x + y)]^2\right\} = E\left\{[\{x - E(x)\} + \{y - E(y)\}]^2\right\}$$
$$= E\left\{[x - E(x)]^2 + [y - E(y)]^2 + 2[x - E(x)][y - E(y)]\right\}$$
$$= V(x) + V(y) + 2C(x, y).$$

If x, y are independent, then $C(x, y) = 0$ and $V(x + y) = V(x) + V(y)$. It is important to note that

> *If x, y are independent, then the covariance is $C(x, y) = 0$. However, the condition $C(x, y) = 0$ does not, in general, imply that x, y are independent.*

A particular case in which $C(x, y) = 0$ does imply the independence of x, y is when both these variables are normally distributed.

The correlation coefficient. To measure the relatedness of x and y, we use the correlation coefficient, defined by

$$\text{Corr}(x, y) = \frac{C(x, y)}{\sqrt{V(x)V(y)}} = \frac{E\{[x - E(x)][y - E(y)]\})}{\sqrt{E\{[x - E(x)]^2\}E\{[y - E(y)]^2\}}}.$$

Notice that this is a number without units.

It can be shown that $-1 \leq \text{Corr}(x, y) \leq 1$. If $\text{Corr}(x, y) = 1$, then there is a perfect positive correlation between x and y, which means that they lie on a straight line of positive slope. If $\text{Corr}(x, y) = -1$, then there is a perfect negative correlation; and the straight line has a negative slope. In other cases, there is a

scatter of points in the plane; and, if $\mathrm{Corr}(x, y)$, then there is no linear relationship between x and y.

These results concerning the range of the correlatison cofficient follow from a version of the Cauchy–Schwarz inequality, which will be established at the end of the next section.

REGRESSION AND CONDITIONAL EXPECTATIONS

Linear conditional expectations. If x, y are correlated, then a knowledge of one of them enables us to make a better prediction of the other. This knowledge can be used in forming conditional expectations.

In some cases, it is reasonable to make the assumption that the conditional expectation $E(y|x)$ is a linear function of x:

$$E(y|x) = \alpha + x\beta. \tag{i}$$

This function is described as a linear regression equation. The error from predicting y by its conditional expectation can be denoted by $\varepsilon = y - E(y|x)$; and therefore we have

$$\begin{aligned} y &= E(y|x) + \varepsilon \\ &= \alpha + x\beta + \varepsilon. \end{aligned}$$

Our object is to express the parameters α and β as functions of the moments of the joint probability distribution of x and y. Usually, the moments of the distribution can be estimated in a straightforward way from a set of observations on x and y. Using the relationship that exits between the parameters and the theoretical moments, we should be able to find estimates for α and β corresponding to the estimated moments.

We begin by multiplying equation (i) throughout by $f(x)$, and by integrating with respect to x. This gives the equation

$$E(y) = \alpha + \beta E(x), \tag{ii}$$

whence

$$\alpha = E(y) - \beta E(x). \tag{iii}$$

These equations shows that the regression line passes through the point $E(x, y) = \{E(x), E(y)\}$ which is the expected value of the joint distribution.

By putting (iii) into (i), we find that

$$E(y|x) = E(y) + \beta\big\{x - E(x)\big\},$$

which shows how the conditional expectation of y differs from the unconditional expectation in proportion to the error of predicting x by taking its expected value.

Now let us multiply (i) by x and $f(x)$ and then integrate with respect to x to provide

$$E(xy) = \alpha E(x) + \beta E(x^2). \tag{iv}$$

Multiplying (ii) by $E(x)$ gives

$$E(x)E(y) = \alpha E(x) + \beta\big\{E(x)\big\}^2, \tag{v}$$

whence, on taking (v) from (iv), we get

$$E(xy) - E(x)E(y) = \beta\Big[E(x^2) - \big\{E(x)\big\}^2\Big],$$

which implies that

$$\begin{aligned}\beta &= \frac{E(xy) - E(x)E(y)}{E(x^2) - \big\{E(x)\big\}^2} \\ &= \frac{E\Big[\big\{x - E(x)\big\}\big\{y - E(y)\big\}\Big]}{E\Big[\big\{x - E(x)\big\}^2\Big]} \\ &= \frac{C(x,y)}{V(x)}.\end{aligned} \tag{vi}$$

Thus, we have expressed α and β in terms of the moments $E(x)$, $E(y)$, $V(x)$ and $C(x,y)$ of the joint distribution of x and y.

It should be recognised that the prediction error $\varepsilon = y - E(y|x) = y - \alpha - x\beta$ is uncorrelated with the variable x. This is shown by writing

$$E\Big[\big\{y - E(y|x)\big\}x\Big] = E(yx) - \alpha E(x) - \beta E(x^2) = 0, \tag{vii}$$

where the final equality comes from (iv). This result is readily intelligible; for, if the prediction error were correlated with the value of x, then we should not be using the information of x efficiently in predicting y.

This section may be concluded by proving a version of the Cauchy–Schwarz inequality that establishes the bounds on $Corr(x,y) = C(x,y)/\sqrt{V(x)V(y)}$, which is the coefficent of the correlation of x and y. Consider the variance of the prediction error

$$E\left(\big[\{y - E(y)\} - \beta\{x - E(x)\}\big]^2\right) = V(y) - 2\beta C(x,y) + \beta^2 V(x) \geq 0.$$

Setting $\beta = C(x,y)/V(x)$ gives

$$V(y) - 2\frac{\{C(x,y)\}^2}{V(x)} + \frac{\{C(x,y)\}^2}{V(x)} \geq 0.$$

whence

$$V(x)V(y) \geq \{C(x,y)\}^2.$$

It follows that $\{Corr(x,y)\}^2 \leq 1$ and, therefore, that $-1 \leq Corr(x,y) \leq 1$.

Empirical Regressions. Imagine that we have a sample of T observations on x and y which are $(x_1, y_1), (x_2, y_2), \ldots, (x_T, y_T)$. Then we can calculate the following empirical or sample moments:

$$\bar{x} = \frac{1}{T}\sum_{t=1}^{T} x_t,$$

$$\bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t,$$

$$S_x^2 = \frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})^2 = \frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})x_t = \frac{1}{T}\sum_{t=1}^{T} x_t^2 - \bar{x}^2,$$

$$S_{xy} = \frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})(y_t - \bar{y}) = \frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})y_t = \frac{1}{T}\sum_{t=1}^{T} x_t y_t - \bar{x}\bar{y},$$

It seems reasonable that, in order to estimate α and β, we should replace the moments in the formulae of (iii) and (vi) by the corresponding sample moments. Thus the estimates of α and β are

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x},$$

$$\hat{\beta} = \frac{\sum(x_t - \bar{x})(y_t - \bar{y})}{\sum(x_t - \bar{x})^2}.$$

The justification of this estimation procedure, which is know as the method of moments, is that, in many of the circumstances under which the sample is liable to be generated, we can expect the sample moments to converge to the true moments of the bivariate distribution, thereby causing the estimates of the parameters to converge likewise to their true values.

Often there is insufficient statistical regularity in the processes generating the variable x to justify our postulating a joint probability density function for x and y. Sometimes the variable is regulated in pursuit of an economic policy in such a way that it cannot be regarded as random in any of the senses accepted by statistical theory. In such cases, we may prefer to derive the estimators of the parameters α and β by methods which make fewer statistical assumptions about x.

When x is a non stochastic variable, the equation

$$y = \alpha + x\beta + \varepsilon$$

is usually regarded as a functional relationship between x and y that is subject to the effects of a random disturbance term ε. It is commonly assumed that, in all instances of this relationship, the disturbance has a zero expected value and a variance which is finite and constant. Thus

$$E(\varepsilon) = 0 \quad \text{and} \quad V(\varepsilon) = E(\varepsilon^2) = \sigma^2.$$

Also it is assumed that the movements in x are unrelated to those of the disturbance term.

The principle of least squares suggests that we should estimate α and β by finding the values which minimise the quantity

$$S = \sum_{t=1}^{T}(y_t - \hat{y}_t)^2$$
$$= \sum_{t=1}^{T}(y_t - \alpha - x_t\beta)^2.$$

This is the sum of squares of the vertical distances—measured parallel to the y-axis—of the data points from an interpolated regression line.

Differentiating the function S with respect to α and setting the results to zero for a minimum gives

$$-2\sum(y_t - \alpha - \beta x_t) = 0, \quad \text{or, equivalently,}$$
$$\bar{y} - \alpha - \beta\bar{x} = 0.$$

This generates the following estimating equation for α:

$$\alpha(\beta) = \bar{y} - \beta\bar{x}. \tag{viii}$$

Next, by differentiating with respect to β and setting the result to zero, we get

$$-2\sum x_t(y_t - \alpha - \beta x_t) = 0. \tag{ix}$$

On substituting for α from (vii) and eliminating the factor -2, this becomes

$$\sum x_t y_t - \sum x_t(\bar{y} - \beta\bar{x}) - \beta\sum x_t^2 = 0,$$

whence we get

$$\hat{\beta} = \frac{\sum x_t y_t - T\bar{x}\bar{y}}{\sum x_t^2 - T\bar{x}^2}$$
$$= \frac{\sum(x_t - \bar{x})(y_t - \bar{y})}{\sum(x_t - \bar{x})^2}.$$

This expression is identical to the one that we have derived by the method of moments. By putting $\hat{\beta}$ into the estimating equation for α under (viii), we derive the same estimate $\hat{\alpha}$ for the intercept parameter as the one obtained by the method of moments.

It is notable that the equation (ix) is the empirical analogue of the equation (vii) which expresses the condition that the prediction error is uncorrelated with the values of x.

The method of least squares does not automatically provide an estimate of $\sigma^2 = E(\varepsilon_t^2)$. To obtain an estimate, we may invoke the method of moments which, in view of the fact that the regression residuals $e_t = y_t - \hat{\alpha} - \hat{\beta}x_t$ represent estimates of the corresponding values of ε_t, suggests an estimator in the form of

$$\tilde{\sigma}^2 = \frac{1}{T}\sum e_t^2.$$

In fact, this is a biased estimator with

$$E\left(T\tilde{\sigma}^2\right) = \left\{T - 2\right\}\sigma^2;$$

so it is common to adopt the unbiased estimator

$$\hat{\sigma}^2 = \frac{\sum e_t^2}{T - 2}.$$