Chapter 3. Multivariate Distributions.

All of the most interesting problems in statistics involve looking at more than a single measurement at a time, at relationships among measurements and comparisons between them. In order to permit us to address such problems, indeed to even formulate them properly, we will need to enlarge our mathematical structure to include *multivariate distributions,* the probability distributions of pairs of random variables, triplets of random variables, and so forth. We will begin with the simplest such situation, that of pairs of random variables or *bivariate distributions,* where we will already encounter most of the key ideas.

3.1 Discrete Bivariate Distributions.

If X and Y are two random variables defined on the same sample space S; that is, defined in reference to the same experiment, so that it is both meaningful and potentially interesting to consider how they may interact or affect one another, we will define their *bivariate probability function* by

$$p(x,y) = P(X = x \quad \text{and} \quad Y = y). \tag{3.1}$$

In a direct analogy to the case of a single random variable (the *univariate* case), $p(x,y)$ may be thought of as describing the distribution of a unit mass in the (x,y) plane, with $p(x,y)$ representing the mass assigned to the point (x,y), considered as a spike at (x,y) of height $p(x,y)$. The total for all possible points must be one:

$$\sum_{\text{all } x} \sum_{\text{all } y} p(x,y) = 1. \tag{3.2}$$

[Figure 3.1]

Example 3.A. Consider the experiment of tossing a fair coin three times, and then, independently of the first coin, tossing a second fair coin three times. Let

$$\begin{aligned} X &= \#\text{Heads for the first coin} \\ Y &= \#\text{Tails for the second coin} \\ Z &= \#\text{Tails for the first coin.} \end{aligned}$$

The two coins are tossed independently, so for any pair of possible values (x,y) of X and Y we have, if $\{X = x\}$ stands for the event "$X = x$",

$$\begin{aligned} p(x,y) &= P(X = x \quad \text{and} \quad Y = y) \\ &= P(\{X = x\} \cap \{Y = y\}) \\ &= P(\{X = x\}) \cdot P(\{Y = y\}) \\ &= P_X(x) \cdot P_Y(y). \end{aligned}$$

On the other hand, X and Z refer to the same coin, and so

$$\begin{aligned} p(x,z) &= P(X = x \quad \text{and} \quad Z = z) \\ &= P(\{X = x\} \cap \{Z = z\}) \\ &= P(\{X = x\}) = p_X(x) \quad \text{if} \quad z = 3 - x \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

This is because we must necessarily have $x + z = 3$, which means $\{X = x\}$ and $\{Z = x - 3\}$ describe the same event. If $z \neq 3 - x$, then $\{X = x\}$ and $\{Z = z\}$ are mutually exclusive and the probability both occur

is zero. These bivariate distributions can be summarized in the form of tables, whose entries are $p(x, y)$ and $p(x, z)$ respectively:

		y			
		0	1	2	3
x	0	$\frac{1}{64}$	$\frac{3}{64}$	$\frac{3}{64}$	$\frac{1}{64}$
	1	$\frac{3}{64}$	$\frac{9}{64}$	$\frac{9}{64}$	$\frac{3}{64}$
	2	$\frac{3}{64}$	$\frac{9}{64}$	$\frac{9}{64}$	$\frac{3}{64}$
	3	$\frac{1}{64}$	$\frac{3}{64}$	$\frac{3}{64}$	$\frac{1}{64}$

$p(x, y)$

		z			
		0	1	2	3
x	0	0	0	0	$\frac{1}{8}$
	1	0	0	$\frac{3}{8}$	0
	2	0	$\frac{3}{8}$	0	0
	3	$\frac{1}{8}$	0	0	0

$p(x, z)$

Now, if we have specified a bivariate probability function such as $p(x, y)$, we can always deduce the respective univariate distributions from it, by addition:

$$p_X(x) = \sum_{\text{all } y} p(x, y), \tag{3.3}$$

$$p_Y(y) = \sum_{\text{all } x} p(x, y), \tag{3.4}$$

The rationale for these formulae is that we can decompose the event $\{X = x\}$ into a collection of smaller sets of outcomes. For example,

$$\{X = x\} = \{X = x \quad \text{and} \quad Y = 0\} \cup \{X = x \quad \text{and} \quad Y = 1\} \cup \cdots$$

$$\cdots \cup \{X = x \quad \text{and} \quad Y = 23\} \cup \cdots$$

where the values of y on the righthand side run through all possible values of Y. But then the events of the righthand side are mutually exclusive (Y cannot have two values at once), so the probability of the righthand side is the sum of the events' probabilities, or $\sum_{\text{all } y} p(x, y)$, while the lefthand side has probability $p_X(x)$.

When we refer to these univariate distributions in a multivariate context, we shall call them the *marginal* probability functions of X and Y. This name comes from the fact that when the addition in (3.3) or (3.4) is performed upon a bivariate distribution $p(x, y)$ written in tabular form, the results are most naturally written in the margins of the table.

Example 3.A (continued). For our coin example, we have the marginal distributions of X, Y, and Z:

		y				
		0	1	2	3	$p_X(x)$
x	0	$\frac{1}{64}$	$\frac{3}{64}$	$\frac{3}{64}$	$\frac{1}{64}$	$\frac{1}{8}$
	1	$\frac{3}{64}$	$\frac{9}{64}$	$\frac{9}{64}$	$\frac{3}{64}$	$\frac{3}{8}$
	2	$\frac{3}{64}$	$\frac{9}{64}$	$\frac{9}{64}$	$\frac{3}{64}$	$\frac{3}{8}$
	3	$\frac{1}{64}$	$\frac{3}{64}$	$\frac{3}{64}$	$\frac{1}{64}$	$\frac{1}{8}$
	$P_y(y)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	

		z				
		0	1	2	3	$p_X(x)$
x	0	0	0	0	$\frac{1}{8}$	$\frac{1}{8}$
	1	0	0	$\frac{3}{8}$	0	$\frac{3}{8}$
	2	0	$\frac{3}{8}$	0	0	$\frac{3}{8}$
	3	$\frac{1}{8}$	0	0	0	$\frac{1}{8}$
	$p_z(z)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	

This example highlights an important fact: you can always find the marginal distributions from the bivariate distribution, but in general you cannot go the other way: you cannot reconstruct the interior of a table (the bivariate distribution) knowing only the marginal totals. In this example, both tables have exactly the same marginal totals, in fact X, Y, and Z all have the same Binomial $\left(3, \frac{1}{2}\right)$ distribution, but

the bivariate distributions are quite different. The marginal distributions $p_X(x)$ and $p_Y(y)$ may describe our uncertainty about the possible values, respectively, of X considered separately, without regard to whether or not Y is even observed, and of Y considered separately, without regard to whether or not X is even observed. But they cannot tell us about the relationship between X and Y, they alone cannot tell us whether X and Y refer to the same coin or to different coins. However, the example also gives a hint as to just what sort of information is needed to build up a bivariate distribution from component parts. In one case the knowledge that the two coins were independent gave us $p(x,y) = p_X(x) \cdot p_Y(y)$; in the other case the complete dependence of Z on X gave us $p(x,z) = p_X(x)$ or 0 as $z = 3 - x$ or not. What was needed was information about how the knowledge of one random variable's outcome may affect the other: *conditional* information. We formalize this as a *conditional probability function,* defined by

$$p(y|x) = P(Y = y|X = x), \tag{3.5}$$

which we read as "the probability that $Y = y$ given that $X = x$." Since "$Y = y$" and "$X = x$" are events, this is just our earlier notion of conditional probability re-expressed for discrete random variables, and from (1.7) we have that

$$\begin{aligned} p(y|x) &= P(Y = y|X = x) \\ &= \frac{P(X = x \quad \text{and} \quad Y = y)}{P(X = x)} \\ &= \frac{p(x,y)}{p_X(x)}, \end{aligned} \tag{3.6}$$

as long as $p_X(x) > 0$, with $p(y|x)$ undefined for any x with $p_X(x) = 0$.

If $p(y|x) = p_Y(y)$ for *all* possible pairs of values (x,y) for which $p(y|x)$ is defined, we say X and Y are *independent variables.* From (3.6), we would equivalently have that X and Y are independent random variables if

$$p(x,y) = p_X(x) \cdot p_Y(y), \quad \text{for all} \quad x, y. \tag{3.7}$$

Thus X and Y are independent only if *all* pairs of events "$X = x$" and "$Y = y$" are independent; if (3.7) should fail to hold for even a single pair (x_o, y_o), X and Y would be *dependent.* In Example 3.A, X and Y are independent, but X and Z are dependent. For example, for $x = 2$, $p(z|x)$ is given by

$$\begin{aligned} p(z|2) &= \frac{p(2,z)}{p_X(2)} \\ &= 1 \quad \text{if } z = 1 \\ &= 0 \quad \text{otherwise,} \end{aligned}$$

so $p(z|x) \neq p_Z(z)$ for $x = 2$, $z = 1$ in particular (and for all other values as well).

By using (3.6) in the form

$$p(x,y) = p_X(x)p(y|x) \quad \text{for all} \quad x, y, \tag{3.8}$$

it is possible to construct a bivariate distribution from two components: either marginal distribution and the conditional distribution of the other variable given the one whose marginal distribution is specified. Thus while marginal distributions are themselves insufficient to build a bivariate distribution, the conditional probability function captures exactly what additional information is needed.

3.2 Continuous Bivariate Distributions.

The distribution of a pair of continuous random variables X and Y defined on the same sample space (that is, in reference to the same experiment) is given formally by an extension of the device used in the univariate case, a density function. If we think of the pair (X, Y) as a random point in the plane, the bivariate probability density function $f(x, y)$ describes a surface in 3-dimensional space, and the probability that (X, Y) falls in a region in the plane is given by the volume over that region and under the surface $f(x, y)$. Since volumes are given as double integrals, the rectangular region with $a < X < b$ and $c < Y < d$ has probability

$$P(a < X < b \quad \text{and} \quad c < Y < d) = \int_c^d \int_a^b f(x, y) dx dy. \tag{3.9}$$

[Figure 3.3]

It will necessarily be true of any bivariate density that

$$f(x, y) \geq 0 \quad \text{for all} \quad x, y \tag{3.10}$$

and

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1, \tag{3.11}$$

that is, the total volume between the surface $f(x, y)$ and the $x - y$ plane is 1. Also, any function $f(x, y)$ satisfying (3.10) and (3.11) describes a continuous bivariate probability distribution.

It can help the intuition to think of a continuous bivariate distribution as a unit mass resting squarely on the plane, not concentrated as spikes at a few separated points, as in the discrete case. It is as if the mass is made of a homogeneous substance, and the function $f(x, y)$ describes the upper surface of the mass.

If we are given a bivariate probability density $f(x, y)$, then we can, as in the discrete case, calculate the *marginal probability densities* of X and of Y; they are given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy \quad \text{for all} \quad x, \tag{3.12}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx \quad \text{for all} \quad y. \tag{3.13}$$

Just as in the discrete case, these give the probability densities of X and Y considered separately, as continuous univariate random variables.

The relationships (3.12) and (3.13) are rather close analogues to the formulae for the discrete case, (3.3) and (3.4). They may be justified as follows: for any $a < b$, the events "$a < X \leq b$" and "$a < X \leq b$ and $-\infty < Y < \infty$" are in fact two ways of describing the same event. The second of these has probability

$$\begin{aligned} \int_{-\infty}^{\infty} \int_a^b f(x, y) dx dy &= \int_a^b \int_{-\infty}^{\infty} f(x, y) dy dx \\ &= \int_a^b \left[\int_{-\infty}^{\infty} f(x, y) dy \right] dx. \end{aligned}$$

We must therefore have

$$P(a < X \leq b) = \int_a^b \left[\int_{-\infty}^{\infty} f(x, y) dy \right] dx \quad \text{for all} \quad a < b,$$

and thus $\int_{-\infty}^{\infty} f(x, y) dy$ fulfills the definition of $f_X(x)$ (given in Section 1.7): it is a function of x that gives the probabilities of intervals as areas, by integration.

In terms of the mass interpretation of bivariate densities, (3.12) amounts to looking at the mass "from the side," in a direction parallel to the y axis. The integral

$$\int_x^{x+dx} \left[\int_{-\infty}^{\infty} f(u,y)dy\right] du \approx \int_{-\infty}^{\infty} f(x,y)dy \cdot dx$$

gives the total mass (for the entire range of y) between x and $x+dx$, and so, just as in the univariate case, the integrand $\int_{-\infty}^{\infty} f(x,y)dy$ gives the density of the mass at x.

Example 3.B. Consider the bivariate density function

$$\begin{aligned} f(x,y) &= y\left(\frac{1}{2}-x\right)+x \quad \text{for} \quad 0<x<1, \quad 0<y<2 \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

One way of visualizing such a function is to look at cross-sections: the function $f\left(x,\frac{1}{2}\right) = \frac{1}{2}\left(\frac{1}{2}-x\right)+x = \frac{x}{2}+\frac{1}{4}$ is the cross-section of the surface, cutting it with a plane at $y=\frac{1}{2}$.

[Figure 3.4]

We can check that $f(x,y)$ is in fact a bivariate density, by evaluating its double integral in an iterated form, taking account of the fact that $f(x,y)$ vanishes outside of the rectangle $0<x<1$, $0<y<2$:

$$\begin{aligned} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy &= \int_0^2\int_0^1 \left[y\left(\frac{1}{2}-x\right)+x\right]dxdy \\ &= \int_0^2 \left\{\int_0^1 \left[y\left(\frac{1}{2}-x\right)+x\right]dx\right\}dy \\ &= \int_0^2 \left[\int_0^1 y\left(\frac{1}{2}-x\right)dx + \int_0^1 xdx\right]dy \\ &= \int_0^2 \left[y\int_0^1 \left(\frac{1}{2}-x\right)dx + \int_0^1 xdx\right]dy \\ &= \int_0^2 \left[y\cdot 0+\frac{1}{2}\right]dy \\ &= \int_0^2 \frac{1}{2}dy = \frac{1}{2}\cdot 2 = 1. \end{aligned}$$

In the process of evaluating this double integral, we found

$$\begin{aligned} \int_{-\infty}^{\infty} f(x,y)dx &= \int_0^1 \left[y\left(\frac{1}{2}-x\right)+x\right]dx \\ &= \frac{1}{2} \quad \text{for all} \quad 0<y<2; \end{aligned}$$

that is, the marginal density of Y is

$$\begin{aligned} f_Y(y) &= \frac{1}{2} \quad \text{for} \quad 0<y<2 \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

We recognize this as the Uniform $(0, 2)$ distribution. We could also calculate

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f(x,y)dy \\ &= \int_0^2 \left[y\left(\frac{1}{2} - x\right) + x \right] dy \quad \text{for} \quad 0 < x < 1 \\ &= \left(\frac{1}{2} - x\right) \int_0^2 y dy + 2x \quad \text{for} \quad 0 < x < 1 \\ &= 2\left(\frac{1}{2} - x\right) + 2x \quad \text{for} \quad 0 < x < 1 \\ &= 1 \quad \text{for} \quad 0 < x < 1 \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

This is just the Uniform $(0, 1)$ distribution! Now if the events "$a < X < b$" and "$c < Y < d$" were independent for all a, b, c, d, we would be able to express $P(a < X < b \text{ and } c < Y < d)$ in terms of these marginal densities:

$$\begin{aligned} P(a < X < b \quad \text{and} \quad c < Y < d) &= P(a < X < b) \cdot P(c < Y < d) \\ &= \int_a^b f_X(x)dx \cdot \int_c^d f_Y(y)dy \\ &= \int_a^b \int_c^d f_X(x) f_Y(y) dy dx, \end{aligned}$$

and thus the product $f_X(x)f_Y(y)$ would fulfil the role of the bivariate density $f(x, y)$. But in this example, $f_X(x)f_Y(y) = \frac{1}{2} \neq f(x, y)$; evidently the events in question are not independent in this example. This example highlights the fact that, just as in the discrete case, we cannot construct a bivariate density from univariate marginal densities, even though we can go the other way and find the marginal densities from the bivariate density. We cannot reconstruct the entire surface $f(x, y)$ from two side views, $f_X(x)$ and $f_Y(y)$. More information is needed; that information is given by the conditional densities.

3.3 Conditional Probability Densities.

In the discrete case we defined the conditional distribution of Y given X as

$$p(y|x) = \frac{p(x,y)}{p_X(x)}, \quad \text{if } p_X(x) > 0. \tag{3.6}$$

We shall follow an analogous course and define the *conditional probability density* of a continuous random variable Y given a continuous random variable X as

$$f(y|x) = \frac{f(x,y)}{f_X(x)}, \quad \text{if } f_X(x) > 0, \tag{3.14}$$

and leave the density $f(y|x)$ undefined if $f_X(x) = 0$. The parallel between (3.6) and (3.14) can be helpful, but it is deceptive. The relationship (3.6) is as we noted, nothing more than a re-expression of our original definition of conditional probability

$$P(E|F) = \frac{P(E \cap F)}{P(F)} \tag{1.7}$$

for the special case $E =$ "$Y = y$" and $F =$ "$X = x$". On the other hand, the definition (3.14) represents a significant extension of (1.7). The reason is that it represents the distribution of Y given $X = x$, but, as

we saw in Section 1.7, for a continuous random variable X we must have $P(X = x) = 0$ for each x, and we have not yet defined conditional probabilities where the given event has probability zero.

The rationale for the definition (3.14) can best be presented by starting with the use to which $f(y|x)$ will be put. For $f(y|x)$ is a density, and it will give conditional probabilities as areas, by integration. In particular, we will have

$$P(a < Y \leq b|X = x) = \int_a^b f(y|x)dy \quad \text{for all } a < b. \tag{3.15}$$

(Indeed, an alternative mathematical approach would start with (3.15) and deduce (3.14)). But what does $P(a < Y \leq b|X = x)$ mean, since $P(X = x) = 0$ and (1.7) does not apply? We proceed heuristically. By $P(a < Y \leq b|X = x)$ we surely must mean something very close to $P(a < Y \leq b|x \leq X \leq x + h)$ for h very small. If $f_X(x) > 0$, this latter probability is well-defined, since $P(x \leq X \leq x + h) > 0$, even though it may be quite small. But then,

$$\begin{aligned} P(a < Y \leq b|X = x) &\approx P(a < Y \leq b|x \leq X \leq x + h) \\ &= \frac{P(x \leq X \leq x + h \quad \text{and} \quad a < Y \leq b)}{P(x \leq X \leq x + h)} \\ &= \frac{\int_a^b \left[\int_x^{x+h} f(u, y)du\right] dy}{\int_x^{x+h} f_X(u)du}. \end{aligned}$$

But if the function $f_X(u)$ does not change value greatly for u in the small interval $[x, x + h]$, we have approximately,

$$\int_x^{x+h} f_X(u)du \approx f_X(x) \cdot h$$

[Figure 3.5].

Similarly, if for fixed y the function $f(u, y)$ does not change value greatly for u in the small interval $[x, x+h]$, we have, again approximately,

$$\int_x^{x+h} f(u, y)du \approx f(x, y) \cdot h.$$

Substituting these approximations into the above expression gives

$$\begin{aligned} P(a < Y \leq b|X = x) &\approx \frac{\int_a^b f(x, y) \cdot hdy}{f_X(x) \cdot h} \\ &= \frac{\int_a^b f(x, y)dy}{f_X(x)} \\ &= \int_a^b \left[\frac{f(x, y)}{f_X(x)}\right] dy. \end{aligned}$$

Comparing this with (3.15) gives us (3.14). This is not a mathematically rigorous demonstration that (3.14) is the proper definition of conditional density, but it captures the essence of one possible rigorous approach, which would combine the passage to the limit of $h \downarrow 0$ with ideas from measure theory.

An intuitive interpretation of conditional densities in terms of the bivariate density is easy: $f(y|x)$ is just the cross-section of the surface $f(x, y)$ at $X = x$, rescaled so that it has total area 1. Indeed, for a fixed x, $f_X(x)$ in the denominator of (3.14) is just the right scaling factor so that the area is 1:

$$\begin{aligned} \int_{-\infty}^{\infty} f(y|x)dy &= \int_{-\infty}^{\infty} \left[\frac{f(x, y)}{f_X(x)}\right] dy \\ &= \frac{1}{f_X(x)} \cdot \int_{-\infty}^{\infty} f(x, y)dy \\ &= \frac{1}{f_X(x)} \cdot f_X(x) \quad \text{from (3.12)} \\ &= 1 \quad \text{for any } x \text{ with } f_X(x) > 0. \end{aligned}$$

Example 3.B (Continued). The conditional density of X given $Y = y$, if $0 < y < 2$, is

$$\begin{aligned} f(x|y) &= \frac{f(x,y)}{f_Y(y)} \\ &= \frac{y\left(\frac{1}{2} - x\right) + x}{\frac{1}{2}} \quad \text{for } 0 < x < 1 \\ &= y(1-2x) + 2x \quad \text{for } 0 < x < 1 \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

This is, except for the scaling factor $1/2$, just the equation for the cross-sections shown in Figure 3.4. For $y = 1/2$, we have

$$\begin{aligned} f\left(x \,\middle|\, \frac{1}{2}\right) &= x + \frac{1}{2} \quad \text{for } 0 < x < 1 \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

We note that these cross-sections, which we now interpret as unscaled conditional densities, vary with y in shape as well as scale. In general, if

$$f(y|x) = f_Y(y) \quad \text{for all } y, \text{ all } x \text{ with } f_X(x) > 0 \tag{3.16}$$

or equivalently

$$f(x|y) = f_X(x) \quad \text{for all } x, \text{ all } y \text{ with } f_Y(y) > 0 \tag{3.17}$$

or equivalently

$$f(x,y) = f_X(x) \cdot f_Y(y) \quad \text{for all } x \text{ and } y \tag{3.18}$$

we shall say that X and Y are *independent random variables.* Any one of these conditions is equivalent to

$$P(a < X < b \quad \text{and} \quad c < Y < d) = P(a < X < b) \cdot P(c < Y < d) \tag{3.19}$$

for all a, b, c, d. If any one of these conditions fails to hold, X and Y are *dependent.*

As we have observed, in Example 3.B X and Y are dependent random variables. In general, by turning (3.14) around, one marginal density (say of X) and one set of conditional densities (say those of Y given X) determine the bivariate density by

$$f(x,y) = f_X(x) f(y|x) \quad \text{for all } x, y. \tag{3.20}$$

Example 3.C. Suppose I choose a number X between 0 and 1 at random, and then you peek at my number and choose a smaller one Y, also at random. If the object of the game is to choose the smallest number, then clearly you will win. But by how much? What is the expectation of your choice, Y? What is the distribution of Y? All of these questions require that we be specific about what "at random" means. Suppose we take it to be uniformly distributed, so I choose according to a Uniform $(0, 1)$ distribution

$$\begin{aligned} f_X(x) &= 1 \quad \text{for } 0 < x < 1 \\ &= 0 \quad \text{otherwise,} \end{aligned}$$

and, conditional on your observing my $X = x$, you choose according to a Uniform $(0, x)$ distribution,

$$\begin{aligned} f(y|x) &= \frac{1}{x} \quad \text{for } 0 < y < x \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

Now from (3.20) we have

$$\begin{aligned} f(x,y) &= f_X(x) f(y|x) \\ &= \frac{1}{x} \quad \text{for } 0 < x < 1 \text{ and } 0 < y < x \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

The bivariate density of X and Y is thus positive over a triangle.

[Figure 3.6]

Once we have found the bivariate density of X and Y, we can proceed with our calculations. The marginal density of Y (that is, the univariate density of Y, ignoring X) is found to be

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f(x,y)dx \\ &= \int_y^1 \frac{1}{x} dx \\ &= \log_e(x) \Big|_y^1 \\ &= -\log_e(y) \quad \text{for } 0 < y < 1 \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

Notice that in performing this calculation it was particularly important to keep track of the range over which the density is positive. For a particular y, $f(x,y) = 1/x$ only for $0 < y < x$ and $0 < x < 1$, so since both of these conditions must be met we have $y < x < 1$ as the range of integration. The marginal density of Y "piles up" mass near 0, reflecting the fact that Y has been selected as smaller than some value of X.

One of the questions asked was, what is the expectation of Y? There is a bit of ambiguity here, does it refer to the marginal distribution of Y, in which the value of X is ignored in making the calculation? Or does it refer to the conditional expectation of Y, given the value of $X = x$? Because no reference is made to "given $X = x$," it is perhaps most natural to calculate the marginal expectation of Y,

$$\begin{aligned} E(Y) &= \int_{-\infty}^{\infty} y f_Y(y) dy \\ &= -\int_0^1 y \log_e(y) dy \\ &= \frac{1}{4} \int_0^{\infty} z e^{-z} dz \quad \text{(by change -of-variable;} \quad z = -2\log_e y, dy = e^{-\frac{z}{2}} dz) \\ &= \frac{1}{4}\Gamma(2) = \frac{1}{4}. \end{aligned}$$

(We will see an easier way of finding $E(Y)$ later.) Thus, since

$$E(X) = \int_0^1 x \cdot 1 dx = \frac{1}{2},$$

the marginal expected value of the smaller number is just half that of X. We can also evaluate the conditional expectation of Y given $X = x$; it is just the expectation of the conditional distribution $f(y|x)$,

$$\begin{aligned} E(Y|X=x) &= \int_0^1 y f(y|x) dy \\ &= \int_0^x y \frac{1}{x} dy \\ &= \frac{1}{x} \cdot \frac{x^2}{2} = \frac{x}{2}, \end{aligned}$$

a result that is obvious since $E(Y|X = x)$ must be the center of gravity of $f(y|x)$ (see Figure 3.6). So either marginally or conditionally, the second choice is expected to be half the first.

What about the other question—what is the expected amount by which Y is less than X, $E(X - Y)$? This requires further development of the mathematics of expectation.

3.4 Expectations of Transformations of Bivariate Random Variables.

A transformation of a bivariate random variable (X, Y) is, as in the univariate case, another random variable that is defined as a function of the pair (X, Y). For example, $h_1(X, Y) = X + Y$, $h_2(X, Y) = X - Y$, and $h_3(X, Y) = X \cdot Y$ are all transformations of the pair (X, Y). We can find the expectation of a transformation $h(X, Y)$ in principle by noting that $Z = h(X, Y)$ is a random variable, and if we can find the distribution of Z, its expectation will follow:

$$E(Z) = \sum_{\text{all z}} z p_Z(z) \quad \text{in the discrete case}$$

or

$$E(Z) = \int_{-\infty}^{\infty} z f_Z(z) dz \quad \text{in the continuous case.}$$

General techniques for finding the distribution of transformations $h(X, Y)$ are generalizations of those used for the univariate case in Section 1.8. But it will turn out that they are unnecesasary in most cases, that we can find $Eh(X, Y)$ without going through the intermediate step of finding the distribution of $h(X, Y)$, just as was the case for univariate transformations (Section 2.2). For by a generalized change-of-variables argument, it can be established that

$$Eh(X, Y) = \sum_{\substack{\text{all } x \\ \text{all } y}} \sum h(x, y) p(x, y) \quad \text{in the discrete case,} \tag{3.21}$$

$$Eh(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y) f(x, y) dx dy \quad \text{in the continuous case.} \tag{3.22}$$

These formulae will hold as long as the expressions they give do not diverge; h need not be monotone or continuous.

Example 3.C (Continued). We had asked earlier for the expectation of $X - Y$, the amount by which the first choice exceeds the second. Here we have $h(X, Y) = X - Y$, and

$$\begin{aligned}
E(X - Y) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - y) f(x, y) dx dy \\
&= \int_0^1 \int_y^1 (x - y) \frac{1}{x} dx dy \\
&= \int_0^1 \left[\int_y^1 x \cdot \frac{1}{x} dx - \int_y^1 y \cdot \frac{1}{x} dx \right] dy \\
&= \int_0^1 \left[\int_y^1 dx - y \int_y^1 \frac{1}{x} dx \right] dy \\
&= \int_0^1 \left[(1 - y) + y \log_e(y) \right] dy \\
&= \int_0^1 (1 - y) dy + \int_0^1 y \log_e(y) dy \\
&= \frac{1}{2} - \frac{1}{4} = \frac{1}{4},
\end{aligned}$$

where we used the fact that we earlier evaluated $-\int_0^1 y \log_e(y) dy = 1/4$. (We may note that here $E(X-Y) = E(X) - E(Y)$; we shall later see that this is always true.)

We could also calculate the expectation of $h(X,Y) = XY$,

$$\begin{aligned} E(XY) &= \int_0^1 \int_y^1 xy \cdot \frac{1}{x} dx dy \\ &= \int_0^1 \left[\int_y^1 y dx \right] dy \\ &= \int_0^1 y \left[\int_y^1 dx \right] dy \\ &= \int_0^1 y(1-y) dy = \frac{1}{6}. \end{aligned}$$

Notice that $E(X)E(Y) = 1/2 \cdot 1/4 = 1/8$; we do *not* have $E(XY) = E(X)E(Y)$. We shall later see that this reflects the fact that X and Y are not independent.

Earlier we calculated $E(Y)$ for this example, a result that depended upon a change-of-variable and recognizing the result as a Gamma function. Here is a simpler alternative route to the same end, based upon treating Y as a special case of a transformation of (X,Y): $h(X,Y) = Y$. Then

$$\begin{aligned} E(Y) &= \int_0^1 \int_y^1 y \cdot \frac{1}{x} dx dy \\ &= \int_0^1 \int_0^x y \cdot \frac{1}{x} dy dx \quad \text{(reversing the order of integration)} \\ &= \int_0^1 \frac{1}{x} \left[\int_0^x y dy \right] dx \\ &= \int_0^1 \frac{1}{x} \cdot \frac{x^2}{2} dx \\ &= \int_0^1 \frac{x}{2} dx = \frac{1}{4}. \end{aligned}$$

(We shall see an even simpler calculation later, in Section 3.10!)

3.5 Mixed Cases.

We have encountered bivariate distributions with both X and Y discrete, and with both X and Y continuous. It is not a long step further to consider situations in which X is discrete and Y is continuous.

Example 3.D. Bayes's Billiard Table. Perhaps the earliest such example was one presented in a famous article published in the *Philosophical Transactions of the Royal Society of London* in 1764, by the Reverend Thomas Bayes. Imagine a flat rectangular table (we might call it a billiard table) with a scale marked on one side, from 0 to 1. A ball is rolled so that it may be considered to be equally likely to come to rest at any point on the table; its final position, Y, is marked on the scale on the side.

[Figure 3.7]

A second ball is then rolled n times, and a count is made of the number of times X it comes to rest to the left of the marked position Y. Then (X,Y) is a bivariate random variable, and X is discrete and Y is continuous.

From the description of the experiment, Y has a Uniform $(0,1)$ distribution, with density

$$\begin{aligned} f_Y(y) &= 1 \quad 0 < y < 1 \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

If we consider each subsequent role as being made independently, then conditional upon $Y = y$, X is a count of the number of successes in n Bernoulli trials where the chance of success on a single trial is y. That is, the conditional distribution of X given $Y = y$ is Binomial (n, y):

$$\begin{aligned} p(x|y) &= \binom{n}{x} y^x (1-y)^{n-x} \quad \text{for } x = 0, 1, \ldots, n \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

For mixed cases as for other cases, we can construct the bivariate distribution by multiplication:

$$\begin{aligned} f(x, y) &= f_Y(y) \cdot p(x|y) \\ &= \binom{n}{x} y^x (1-y)^{n-x} \quad \text{for } 0 \le y \le 1 \quad \text{and} \quad x = 0, 1, 2, \ldots, n, \\ &= 0 \quad \text{otherwise.} \end{aligned} \tag{3.23}$$

This distribution is a hybrid; it is discrete in the argument x and continuous in the argument y. It might be pictured as a series of parallel sheets resting on the $x - y$ plane, concentrated on the lines $x = 0, x = 1, \ldots, x = n$. For example, the cross-section at $x = 1$ is the sheet whose upper edge has the outline described, as a function of y, by

$$f(1, y) = ny(1-y)^{n-1} \quad 0 \le y \le 1$$

[Figure 3.8]

The cross-section parallel to the x-axis at a particular value of y shows the spikes of the Binomial (n, y) distribution.

For dealing with mixed cases, we can use our results for the discrete and continuous cases. For example, the (discrete) marginal distribution of X is

$$\begin{aligned} p_X(x) &= \int_{-\infty}^{\infty} f(x, y) dy \\ &= \int_0^1 \binom{n}{x} y^x (1-y)^{n-x} dy \\ &= \binom{n}{x} \int_0^1 y^x (1-y)^{n-x} dy \\ &= \binom{n}{x} \cdot B(x+1, n-x+1) \quad \text{(from (2.12))} \\ &= \binom{n}{x} \cdot \frac{1}{\binom{n}{x}(n+1)} \\ &= \frac{1}{n+1} \quad \text{for } x = 0, 1, 2, \ldots, n, \\ &= 0 \quad \text{otherwise.} \end{aligned} \tag{3.24}$$

(We evaluated the integral by recognizing it was a Beta function (2.12), and by using the identity given just before (2.12), where the reciprocal of the Beta function is given for integer arguments.) We recognize the result as the discrete uniform distribution of Example 2.0. An observer of this experiment who was not in a position to note the value of Y, but would only learn the count X, would find all possible values of X equally likely.

We can also calculate the conditional distribution of Y given $X = x$. For $x = 0, 1, \ldots, n$ it will be a density, given by

$$f(y|x) = \frac{f(x, y)}{p_X(x)} \tag{3.25}$$

$$= \frac{\binom{n}{x} y^x (1-y)^{n-x}}{\left(\frac{1}{n+1}\right)}$$
$$= (n+1)\binom{n}{x} y^x (1-y)^{n-x} \quad \text{for } 0 \le y \le 1$$
$$= 0 \quad \text{otherwise.}$$

(If x is not an integer, $p_X(x) = 0$ and $f(y|x)$ is undefined.) We can recognize this as a Beta density, with $\alpha = x+1$ and $\beta = n-x+1$. We shall discuss the potential uses of these calculations for statistical inference in the next chapter.

3.6 Higher Dimensions.

We have concentrated on bivariate distributions because of their mathematical simplicity, and because they are easier to visualize than higher dimensional distributions are. The same ideas carry over to higher dimensions, however. If X_1, X_2, and X_3 are three discrete random variables, their trivariate distribution is described by a probability function

$$p(x_1, x_2, x_3) = P(X_1 = x_1 \quad \text{and} \quad X_2 = x_2 \quad \text{and} \quad X_3 = x_3),$$

and

$$\sum_{\text{all } x_1, x_2, x_3} p(x_1, x_2, x_3) = 1.$$

If X_1, X_2, and X_3 are continuous, we would describe their distribution by a density $f(x_1, x_2, x_3)$, where

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1, x_2, x_3)dx_1dx_2dx_3 = 1,$$

and the same integral over a restricted range in X_1, X_2, X_3, space would give the probability the random point (X_1, X_2, X_3), falls in that restricted range. More general multivariate distributions

$$p(x_1, x_2, \ldots, x_n)$$

and

$$f(x_1, x_2, \ldots, x_n)$$

would be defined analogously.

We will say a collection of random variables are *independent* if their multivariate distribution factors into a product of the univariate distributions for all values of the arguments: in the discrete case,

$$p(x_1, x_2, \ldots, x_n) = p_{X_1}(x_1)p_{X_2}(x_2)\cdots p_{X_n}(x_n); \tag{3.26}$$

in the continuous case,

$$f(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1)f_{X_2}(x_2)\cdots f_{X_n}(x_n). \tag{3.27}$$

The terms on the right here are the *marginal distributions* of the X_i's. They can be found here, as in the bivariate case, by summation or integration. For example, in the trivariate case

$$p_{X_1}(x_1) = \sum_{\text{all } x_2, x_3}\sum p(x_1, x_2, x_3);$$

or

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1, x_2, x_3)dx_2dx_3.$$

We may also have multivariate marginal distributions: If X_1, X_2, X_3, and X_4 have a continuous four dimensional distribution, the marginal density of (X_1, X_2) is

$$f(x_1, x_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2, x_3, x_4) dx_3 dx_4.$$

Conditional distributions can be found in a manner analogous to that in the bivariate case. For example, the conditional bivariate density of X_3 and X_4 given $X_1 = x_1$ and $X_2 = x_2$ is

$$f(x_3, x_4 | x_1, x_2) = \frac{f(x_1, x_2, x_3, x_4)}{f(x_1, x_2)}.$$

Expectations of transformations are found in the same way: in the discrete case,

$$E(h(X_1, X_2, \ldots, X_n)) = \sum_{\text{all } x\text{'s}} h(x_1, dx_2, \ldots, x_n) p(x_1, x_2, \ldots, x_n), \tag{3.28}$$

in the continuous case,

$$E(h(X_1, X_2, \ldots, X_n)) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} h(x_1, \ldots, x_n) f(x_1, \ldots, x_n) dx_1, \ldots, dx_n. \tag{3.29}$$

Mathematically, the problems that arise in higher dimensions can be quite intricate, though they will be considerably simplified in many of the cases we will consider by the fact that the random variables will be specified to be independent.

3.7 Measuring Multivariate Distributions.

Describing multivariate distributions can be a complicated task. The entire distribution can be given by the probability function or probability density function, but such a description can be hard to visualize. Univariate distributions can often be better understood by graphing them, but this becomes more difficult with multivariate distributions: the distribution of a trivariate random variable requires a four dimensional picture. Graphs of marginal distributions can help, but as we have emphasized, it is only under additional assumptions (such as independence) that the marginal distributions tell the whole story.

In the bivariate case, we could use, for example, the fact that

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

and graph $f(y|x)$ as a function of y for a selection of values of x, together with a graph of $f_X(x)$. This would effectively show the surface by showing cross-sections (the $f(y|x)$'s) and a side view ($f_X(x)$). In the bivariate case we could also show the surface by plotting level curves; that is, through contour plots of the type used in geography to show three dimensions in two.

[Figure 3.9]

This is tantamount to slicing the surface $f(x, y)$ parallel to the $x - y$ plane at various heights and graphing the cross-sections on one set of axes.

Even though with special computer technology these graphical devices are becoming increasingly popular, and some can in limited ways be adapted to higher dimensions, we will have frequent need for numerical summary measures which, while they do not describe the entire distribution, at least capture some statistically important aspects of it. The measures we consider here are based upon expectations of transformations of the multivariate random variable, such as $Eh(X, Y)$.

A natural starting point for multivariate summary measures is linear transformations; in the bivariate case,

$$h(X, Y) = aX + bY, \quad \text{where } a \text{ and } b \text{ are constants.}$$

The following theorem shows that the expectations of linear transformations are particularly easy to evaluate:

Theorem: For any constants a and b and any multivariate random variable (X, Y),

$$E(aX + bY) = aE(X) + bE(Y), \tag{3.30}$$

as long as all of these expressions are well defined.

Proof (for the continuous case). From (3.22) we have

$$\begin{aligned}
E(aX + bY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (ax + by) f(x, y) dx dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [axf(x, y) + byf(x, y)] dx dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} axf(x, y) dx dy + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} byf(x, y) dx dy \\
&= a\int_{-\infty}^{\infty} x\left[\int_{-\infty}^{\infty} f(x, y) dy\right] dx + b\int_{-\infty}^{\infty} y\left[\int_{-\infty}^{\infty} f(x, y) dx\right] dy \\
&= a\int_{-\infty}^{\infty} xf_X(x) dx + b\int_{-\infty}^{\infty} yf_Y(y) dy \quad \text{(using (3.12), (3.13))} \\
&= aE(X) + bE(Y).
\end{aligned}$$

(Alternatively, we could have skipped the fourth and fifth steps here by noticing, for example, that (3.22) with $h(x, y) = ax$ tells us that

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} axf(x, y) dx dy = E(aX) = aE(X).)$$

The proof for the discrete case is similar.

These results can be immediately generalized by induction: For any $a_1, a_2, \ldots, a_n$ and any multivariate random variable $(X_1, X_2, \ldots, X_n)$,

$$E\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i E(X_i), \tag{3.31}$$

as long as all expressions are well-defined (that is, the expectations do not diverge.)

This Theorem can be quite helpful; it tells us that it is easy to compute expectations of linear transformations such as $aX + bY$, because all we need to know are the separate marginal distributions of X and Y. But the Theorem has a negative aspect as well: since $E(aX + bY)$ can be determined from facts about the marginal distributions alone for any a and b, in particular from knowing just $E(X)$ and $E(Y)$, and since the marginal distributions do not allow us to reconstruct the bivariate distribution, it is hopeless to try to capture information about how X and Y vary together by looking at expectations of linear transformations! It is simply impossible for summary measures of the form $E(aX + bY)$ to tell us about the degree of dependence (or lack of it) between X and Y. To learn about how X and Y vary together, we must go beyond expectations of linear transformations.

3.8 Covariance and Correlation.

The simplest nonlinear transformation of X and Y is XY, and we could consider $E(XY)$ as a summary. Because $E(XY)$ is affected by where the distributions are centered as well as how X and Y vary together, we start instead with the product of $X - E(X)$ and $Y - E(Y)$, that is the expectation of

$$h(X,Y) = (X - \mu_X)(Y - \mu_Y).$$

This is called the *covariance* of X and Y, denoted

$$\text{Cov}(X,Y) = E[(X - \mu_X)(Y - \mu_Y)]. \tag{3.32}$$

By expanding

$$h(X,Y) = XY - \mu_X Y - X\mu_Y + \mu_X\mu_Y$$

and using (3.31) we arrive at an alternative expression,

$$\begin{aligned}\text{Cov}(X,Y) &= E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X\mu_Y \\ &= E(XY) - \mu_X\mu_Y - \mu_Y\mu_X + \mu_X\mu_Y\end{aligned}$$

or

$$\text{Cov}(X,Y) = E(XY) - \mu_X\mu_Y. \tag{3.33}$$

It is immediate from either (3.32) or (3.33) that $\text{Cov}(X,Y) = \text{Cov}(Y,X)$.

Example 3.C. (Continued) For the density

$$\begin{aligned}f(x,y) &= \frac{1}{x} \quad \text{for } 0 < y < x < 1 \\ &= 0 \quad \text{otherwise,}\end{aligned}$$

we found

$$E(X) = \mu_X = \frac{1}{2},$$

$$E(Y) = \mu_Y = \frac{1}{4},$$

and

$$E(XY) = \frac{1}{6}.$$

We then have

$$\text{Cov}(X,Y) = \frac{1}{6} - \frac{1}{2}\cdot\frac{1}{4} = \frac{1}{24}.$$

Note that in view of (3.33), Covariance may be considered as a measure of exactly how much $E(XY)$ differs from $E(X)E(Y)$. But how are we to interpret the magnitude of this "measure"? To begin with and to help see how it relates qualitatively to the bivariate distribution of X and Y, consider Figure 3.10, showing contours for a bivariate distribution. When

[Figure 3.10]

X and Y are in quadrant I or quadrant III, $(X - \mu_x)(Y - \mu_Y)$ is positive; in quadrants II and IV $(X - \mu_X)(Y - \mu_Y)$ is negative. If there is a tendency for (X,Y) to be in I or III, as in Figure 3.10, we would expect positive values to dominate and $\text{Cov}(X,Y) > 0$; covariance is, after all, a weighted average value of $(X - \mu_X)(Y - \mu_Y)$, weighted by the probabilities of the different values of this product.

Qualitatively, a positive $\mathrm{Cov}(X,Y)$ suggests that large values of X are associated with large values of Y, and small values of X are associated with small values of Y. At one extreme, if X and Y are the same random variable, we have

$$\begin{aligned}\mathrm{Cov}(X,X) &= E(X\cdot X) - \mu_X\mu_X \\ &= \mathrm{Var}(X).\end{aligned} \tag{3.34}$$

But covariance can be negative as well, if large values of X are associated with small values of Y. The extreme case is $Y=-X$, in which case, since $\mu_{-X}=-\mu_X$,

$$\begin{aligned}\mathrm{Cov}(X,-X) &= E(X(-X)) - \mu_X\mu_{-X} \\ &= -E(X^2) + \mu_X^2 = -\mathrm{Var}(X).\end{aligned} \tag{3.35}$$

We can think of $X=Y$ and $X=-Y$ as representing extremes of positive and negative dependence, and intermediate between these extremes is the case of independence. If X and Y are independent,

$$\mathrm{Cov}(X,Y)=0. \tag{3.36}$$

To see this, consider

$$\begin{aligned}E(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_X(x)f_Y(y)dxdy \quad \text{by independence} \\ &= \int_{-\infty}^{\infty} xf_X(x)dx \cdot \int_{-\infty}^{\infty} yf_Y(y)dy \\ &= \mu_X\mu_Y;\end{aligned} \tag{3.37}$$

a similar argument gives $E(XY)=\mu_X\mu_Y$ in the discrete case.

But $\mathrm{Cov}(X,Y)=0$ is not sufficient to guarantee independence; intuitively there can be an exact balance between the weighted values of $h(X,Y)=(X-\mu_X)(Y-\mu_Y)$ in quadrants I and III and in quadrants II and IV of Figure 3.10 for dependent X and Y.

Example 3.E. Suppose (X,Y) has the bivariate discrete distribution $p(x,y)$ given by the table

		y				$p_X(x)$
		-2	-1	1	2	
	-2	$\frac{1}{10}$	0	0	0	$\frac{1}{10}$
	-1	0	0	$\frac{2}{5}$	0	$\frac{2}{5}$
x	1	0	$\frac{2}{5}$	0	0	$\frac{2}{5}$
	2	0	0	0	$\frac{1}{10}$	$\frac{1}{10}$
	$p_X(y)$	$\frac{1}{10}$	$\frac{2}{5}$	$\frac{2}{5}$	$\frac{1}{10}$	

Then

$$\begin{aligned}\mu_X &= \mu_Y = 0 \\ E(XY) &= (-2)(-2)\frac{1}{10} + (-1)\cdot 1\cdot\frac{2}{5} + 1\cdot(-1)\frac{2}{5} + 2\cdot 2\cdot\frac{1}{10} \\ &= 0,\end{aligned}$$

and so

$$\mathrm{Cov}(X,Y)=0.$$

Yet $p(x,y) \neq p_X(x)p_Y(y)$ for any possible pair (x,y); X and Y are dependent.

[Figure 3.11]

In general, we shall refer to two random variables with $\mathrm{Cov}(X,Y) = 0$ as *uncorrelated,* keeping in mind that they may still be dependent, though the dependence will not be of the monotone sort exhibited in Figure 3.10.

The most important single general use of covariance, and the best means of interpreting it quantitatively, is by viewing it as a correction term that arises in calculating the variance of sums. By using (3.30) with $a = b = 1$, we have already that

$$E(X+Y) = E(X) + E(Y) \tag{3.38}$$

(or $\mu_{X+Y} = \mu_X + \mu_Y$). Now the variance $\mathrm{Var}(X+Y)$ is the expectation of

$$\begin{aligned}(X+Y-\mu_{X+Y})^2 &= [X+Y-(\mu_X+\mu_Y)]^2 \quad \text{by (3.38)} \\ &= [(X-\mu_X)+(Y-\mu_Y)]^2 \\ &= (X-\mu_X)^2 + (Y-\mu_Y)^2 + 2(X-\mu_X)(Y-\mu_Y).\end{aligned}$$

Now, using (3.31) with $X_1 = (X-\mu_X)^2$, $X_2 = (Y-\mu_Y)^2$, $X_3 = (X-\mu_X)(Y-\mu_Y)$, $a_1 = a_2 = 1$, and $a_3 = 2$, we have

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y). \tag{3.39}$$

This is true as long as all expressions are well-defined (ie. do not diverge).

Comparing (3.38) and (3.39) illustrates the point made earlier: $E(X+Y)$ depends only on the marginal distributions, but $\mathrm{Var}(X+Y)$ reflects the bivariate behavior through the covariance correction term. In the extreme cases $X = Y$ and $X = -Y$ we have

$$\begin{aligned}\mathrm{Var}(X+X) &= \mathrm{Var}(X) + \mathrm{Var}(X) + 2\,\mathrm{Cov}(X,X) \\ &= 4\,\mathrm{Var}(X) \quad \text{(using 3.34)},\end{aligned}$$

while

$$\begin{aligned}\mathrm{Var}(X+(-X)) &= \mathrm{Var}(X) + \mathrm{Var}(-X) + 2\,\mathrm{Cov}(X,-X) \\ &= 0,\end{aligned}$$

from (3.35). The first of these agrees with (2.32) with $a = 2$, $b = 0$; the second is obvious since $X+(-X) = 0$ has zero variance. In the special case where X and Y are independent (or even uncorrelated), $\mathrm{Cov}(X,Y) = 0$ and

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y). \tag{3.40}$$

Thus $\mathrm{Cov}(X,Y)$ can be interpreted quantitatively; it is half the correction term that should be applied to the formula for $\mathrm{Var}(X+Y)$ for the independent case.

The interpretation of covariance as a correction factor for variances of sums gives meaning to it as a magnitude, but it reveals a shortcoming to $\mathrm{Cov}(X,Y)$ as a measure of dependence: $\mathrm{Cov}(X,Y)$ changes with multiplicative changes in the scales of measurement of X and Y. In fact, since

$$aX + b - \mu_{aX+b} = a(X-\mu_X)$$

and

$$cY + d - \mu_{cY+d} = c(Y-\mu_Y),$$

we have

$$\mathrm{Cov}(aX+b, cY+d) = ac\,\mathrm{Cov}(X,Y) \tag{3.41}$$

for all constants a, b, c, d. Large covariance may be due to a high degree of association or dependence between X and Y; it may also be due to the choice of scales of measurement. To eliminate this scale effect we define the *correlation* of X and Y to be the covariance of the standardized X and Y,

$$\begin{aligned}\rho_{XY} &= E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right] \\ &= \mathrm{Cov}\left(\frac{X-\mu_X}{\sigma_X}, \frac{Y-\mu_Y}{\sigma_Y}\right) \\ &= \frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y}.\end{aligned} \tag{3.42}$$

Example 3.C (Continued) For the bivariate density

$$\begin{aligned}f(x,y) &= \frac{1}{x} \quad \text{for } 0<y<x<1 \\ &= 0 \quad \text{otherwise,}\end{aligned}$$

we found $E(X) = \frac{1}{2}$, $E(Y) = \frac{1}{4}$, and $\mathrm{Cov}(X,Y) = \frac{1}{24}$. Furthermore,

$$E(X^2) = \int_0^1 x^2 \cdot 1 dx = \frac{1}{3}$$

and

$$\begin{aligned}E(Y^2) &= \int_0^1 \int_0^x y^2 \cdot \frac{1}{x}\, dydx \\ &= \int_0^1 \frac{1}{x} \cdot \frac{x^2}{3}\, dx = \int_0^1 \frac{x^2}{3}\, dx = \frac{1}{9},\end{aligned}$$

so

$$\mathrm{Var}(X) = \frac{1}{3} - \left(\frac{1}{2}\right)^2 = \frac{1}{12}$$

$$\mathrm{Var}(Y) = \frac{1}{9} - \left(\frac{1}{4}\right)^2 = \frac{7}{144}.$$

We then have

$$\rho_{XY} = \frac{\frac{1}{24}}{\sqrt{\frac{1}{12}\cdot\frac{7}{144}}} = \sqrt{\frac{3}{7}} = .65.$$

If X and Y were independent with the same marginal distributions, we would have

$$\mathrm{Var}(X+Y) = \frac{1}{12} + \frac{7}{144} = \frac{19}{144}.$$

Since they are positively correlated, the variance of $X+Y$ will exceed this by

$$2\,\mathrm{Cov}(X,Y) = \frac{1}{12} = \frac{12}{144},$$

or

$$\mathrm{Var}(X+Y) = \frac{31}{144}.$$

It follows immediately from (3.41) and (2.32) that as long as $ac > 0$,

$$\rho_{aX+b,cY+d} = \rho_{X,Y}, \tag{3.43}$$

while if $ac < 0$,

$$\rho_{aX+b,cY+d} = -\rho_{XY}. \tag{3.44}$$

That is, correlation is a "scale-free" measure of the bivariate distribution of X and Y. It reflects some aspects of the dependence of X and Y. If X and Y are independent,

$$\rho_{XY} = 0, \tag{3.45}$$

but if X and Y are uncorrelated, that is, if $\rho_{XY} = 0$, it does not follow that X and Y are independent (as Example 3.E shows).

Correlation may be interpreted as a correction term for the variance of the sum of the standardized variables

$$W = \frac{X - \mu_X}{\sigma_X}, V = \frac{Y - \mu_Y}{\sigma_Y},$$

and looking at it from this point of view will tell us its range of possible values. We have

$$\begin{aligned} \mathrm{Var}(W + V) &= \mathrm{Var}(W) + \mathrm{Var}(V) + 2\rho_{XY} \\ &= 1 + 1 + 2\rho_{XY} \\ &= 2(1 + \rho_{XY}). \end{aligned}$$

Since we must have $\mathrm{Var}(W + V) \geq 0$ for any bivariate distribution, it is always true that $1 + \rho_{XY} \geq 0$, or $\rho_{XY} \geq -1$. Since this must also apply with $-Y$ for Y, we also have $\rho_{X,-Y} \geq -1$ and (3.44) implies

$$\rho_{X,-Y} = -\rho_{X,Y}, \tag{3.46}$$

so we must also have $-\rho_{XY} \geq -1$ or $\rho_{XY} \leq 1$. Together, these give us the range of possible values of ρXY as

$$-1 \leq \rho_{XY} \leq 1. \tag{3.47}$$

We have already encountered the extremes: $\rho_{XX} = 1$, while $\rho_{X,-X} = -1$.

Even though correlation is "scale-free," it is extremely limited as a general measure of dependence or association. The existence of examples such as Example 3.E suggest what is true, that no single number can adequately capture all the potential complexities of a bivariate distribution. We shall see, however, that in the more restricted setting of a multivariate normal distribution, correlation measures the degree of dependence perfectly.

3.9 Covariance in Higher Dimensions.

For general bivariate distributions, covariance and correlation are most useful as correction factors for variances of sums, and of limited usefulness as measures of dependence. As we would expect would be the case, measuring dependence in higher dimensions is even more difficult. However, for the limited purpose of calculating variances of sums, bivariate covariance provides the needed correction factor, even in higher dimensions.

Suppose $(X_1, \ldots, X_n)$ is an n dimensional random variable. Then from (3.31),

$$\mu_{X_1+\cdots+X_n} = \mu_{\Sigma X_i} = \sum_{i=1}^{n} \mu_{X_i}, \tag{3.48}$$

and

$$\begin{aligned} \left(\sum_{i=1}^{n} X_i - \mu_{\Sigma X_i}\right)^2 &= \left(\sum_{i=1}^{n} (X_i - \mu_{X_i})\right)^2 \\ &= \sum_{i=1}^{n} (X_i - \mu_{X_i})^2 + \sum_{\substack{\text{all } i,j \\ i \neq j}} \sum (X_i - \mu_{X_i})(X_j - \mu_{X_j}) \end{aligned}$$

Taking expectations (again using 3.31),

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i) + \sum_{\substack{\text{all } i,j \\ i \neq j}} \sum \text{Cov}(X_i, X_j). \tag{3.49}$$

Slightly more generally, we also have

$$\text{Var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \,\text{Var}(X_i) + \sum_{\substack{\text{all } i,j \\ i \neq j}} \sum a_i a_j \,\text{Cov}(X_i, X_j), \tag{3.50}$$

for any constants $a_1, a_2, \ldots, a_n$. In the special case where the X's are mutually independent (or, even if they are pairwise uncorrelated), the last double sum in each case equals zero, and we have

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i). \tag{3.51}$$

$$\text{Var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \,\text{Var}(X_i). \tag{3.52}$$

Thus the double sum in (3.49) and (3.50) may be considered as a correction term, capturing the total effect of the dependence among the X's as far as the variance of sums is concerned.

We shall be particularly interested in the behavior of arithmetic averages of independent random variables $X_1, \ldots, X_n$, denoted

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i. \tag{3.53}$$

From (3.31) and (3.52), with $a_1 = \cdots = a_n = 1/n$,

$$E(\bar{X}) = \frac{1}{n} \sum_{i=1}^{n} \mu_{X_i}, \tag{3.54}$$

$$\text{Var}(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^{n} \sigma_{X_i}^2. \tag{3.55}$$

If the X's have the same expectations and variances ($\mu_{X_1} = \cdots = \mu_{X_n} = \mu$; $\sigma_{X_1{}^2} = \cdots = \sigma_{X_n}^2 = \sigma^2$), we get

$$E(\bar{X}) = \mu, \tag{3.56}$$

$$\text{Var}(\bar{X}) = \sigma^2/n. \tag{3.57}$$

For many statistical purposes it is extraordinarily useful to represent higher dimensional random variables in terms of vector-matrix notation. We can write the multivariate random variable $(X_1, X_2, \ldots, X_n)$ as a column vector, denoted

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \cdot \\ \cdot \\ \cdot \\ X_n \end{pmatrix},$$

and the expectations of the X_i's as another column vector:

$$\boldsymbol{\mu}_{\mathbf{X}} = \begin{pmatrix} \mu_{X_1} \\ \mu_{X_2} \\ \cdot \\ \cdot \\ \cdot \\ \mu_{X_n} \end{pmatrix}.$$

Symbolically, we then have

$$E(\mathbf{X}) = \boldsymbol{\mu}_{\mathbf{X}};$$

that is, the expectation of a random vector is the vector of expectations of the components.

If $\mathbf{a}$ is a column vector of constants,

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \cdot \\ \cdot \\ \cdot \\ a_n \end{pmatrix},$$

then (3.31) becomes

$$E(\mathbf{a}^{\mathbf{T}}\mathbf{X}) = \mathbf{a}^{\mathbf{T}} E(\mathbf{X}) = \mathbf{a}^{\mathbf{T}} \boldsymbol{\mu}_{\mathbf{X}}, \tag{3.58}$$

where $\mathbf{a}^{\mathbf{T}}$ is the transpose of $\mathbf{a}$, the row vector

$$\mathbf{a}^{\mathbf{T}} = (a_1 a_2 \cdots a_n),$$

and $\mathbf{a}^{\mathbf{T}}\mathbf{X} = \sum_{i=1}^{n} a_i X_i$ is the usual matrix product. It is customary to arrange the variances and covariances of the X_i in the form of a square $n \times n$ matrix called the *Covariance matrix,*

$$\mathrm{Cov}(\mathbf{X}) = \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \mathrm{Cov}(X_1, X_3) & \cdots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_1, X_2) & \mathrm{Var}(X_2) & \mathrm{Cov}(X_2 X_3) & \cdots & \mathrm{Cov}(X_2 X_n) \\ \mathrm{Cov}(X_1, X_3) & \cdots & & & \\ \vdots & & \ddots & & \vdots \\ \mathrm{Cov}(X_1 X_n) & \cdots & \cdots & & \mathrm{Var}(X_n) \end{pmatrix} \tag{3.59}$$

Since $\mathrm{Var}(X_i) = \mathrm{Cov}(X_i, X_i)$, the entry in row i and column j is just $\mathrm{Cov}(X_i, X_j)$. The reason for this arrangement is that it permits many statistical operations to be written quite succinctly. For example, (3.52) becomes

$$\mathrm{Var}(\mathbf{a}^{\mathbf{T}}\mathbf{X}) = \mathbf{a}^{\mathbf{T}} \,\mathrm{Cov}(\mathbf{X})\mathbf{a}. \tag{3.60}$$

3.10 Conditional Expectation.

If (X, Y) is a bivariate random variable, the entire bivariate distribution can be reconstructed from one marginal distribution, say $f_X(x)$, and the other conditional distributions, say $f(y|x)$, for all x. Now the conditional distributions $f(y|x)$ are a family of distributions, a possibly different distribution for the variable Y for every value of x. The manner in which these distributions vary with x is informative about how X and Y are related, and the distributions $f(y|x)$ can be described in part, just as for any univariate distributions, by their expectations and variances. This does not involve the introduction of a new concept, only a new use for our previous concept. Thus we have the *conditional expectation* of Y given $X = x$ defined by, in the discrete case

$$E(Y|X = x) = \sum_{\text{all } y} yp(y|x), \tag{3.61}$$

in the continuous case

$$E(Y|X = x) = \int_{-\infty}^{\infty} yf(y|x)dy. \tag{3.62}$$

Other expectations can be found by application of results previously discussed; for example, for the continuous case

$$E(h(Y)|X = x) = \int_{-\infty}^{\infty} h(y)f(y|x)dy. \tag{3.62}$$

This can be applied to define the *conditional variance* of Y given $X = x$ by, in the discrete case,

$$\text{Var}(Y|X = x) = \sum_{\text{all } y} (y - E(Y|X = x))^2 p(y|x), \tag{3.63}$$

and in the continuous case,

$$\text{Var}(Y|X = x) = \int_{-\infty}^{\infty} (y - E(Y|X = x))^2 f(y|x)dy. \tag{3.64}$$

Note that $\text{Var}(Y|X = x) = E(Y^2|X = x) - [E(Y|X = x)]^2$ follows from (2.26).

Now, as emphasized above, $E(Y|X = x)$ and $\text{Var}(Y|X = x)$ are, despite the new names, only ordinary expectations and variances for the univariate distributions $p(y|x)$ or $f(y|x)$. However, there are many of them, one $E(Y|X = x)$ and one $\text{Var}(Y|X = x)$ for each possible value of X. Indeed, looked at this way, they are transformations of X, just as $h(X) = X^2$ or $h(X) = 3X + 2$ are transformations of X. Thus if we consider X as a random variable, these particular transformations of X may be considered as random variables. Viewed this way, we will write $E(Y|X)$ and mean the random variable which, when $X = x$, takes the value $E(Y|X = x)$. Similarly, $\text{Var}(Y|X)$ is a random variable that equals $\text{Var}(Y|X = x)$ when $X = x$. Furthermore, as the random variables with univariate probability distributions, we can discuss the expectation and variance of $E(Y|X)$ and $\text{Var}(Y|X)$. Now, in practice it may be quite difficult to actually determine the distributions of $E(Y|X)$ and $\text{Var}(Y|X)$, but two identities involving their expectations, play a remarkable role in statistical theory, as we shall see later. These are given in the following theorem.

Theorem: For any bivariate random variable (X, Y),

$$E[E(Y|X)] = E(Y), \tag{3.65}$$

$$\text{Var}(Y) = E(\text{Var}(Y|X)) + \text{Var}(E(Y|X)). \tag{3.66}$$

(The first of these has been called "Adam's Theorem," because for every claimed first use, someone has found an earlier use.)

Proof: (for the continuous case). The expression "$E[E(Y|X)]$" simply means $E(h(X))$, for the $h(x)$ given by (3.62). We have

$$\begin{aligned} E[E(Y|X)] &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} y f(y|x) dy \right] f_X(x) dx \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f(y|x) f_X(x) dy dx \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f(x,y) dy dx \\ &= E(Y), \end{aligned}$$

using (3.22) with $h(x,y) = y$. For the second result,

$$\begin{aligned} E(\mathrm{Var}(Y|X)) &= E\{E(Y^2|X) - [E(Y|X)]^2\} \\ &= E[E(Y^2|X)] - E[E(Y|X)]^2 \\ &= E(Y^2) - E[E(Y|X)]^2, \end{aligned}$$

using (3.65) with Y^2 for Y. Also,

$$\begin{aligned} \mathrm{Var}(E(Y|X)) &= E[E(Y|X)]^2 - \{E[E(Y|X)]\}^2 \\ &= E[E(Y|X)]^2 - [E(Y)]^2 \end{aligned}$$

using (3.65) directly. Adding these two expressions gives the result (3.66), since two terms cancel to give $E(Y^2) - [E(Y)]^2 = \mathrm{Var}(Y)$.

Example 3.C (Continued). Here the distribution of X is Uniform $(0,1)$ and that of Y given $X = x$ is Uniform $(0,x)$. Now it is immediate that the expectation of a Uniform random variable is just the midpoint of its range:

$$E(Y|X = x) = \frac{x}{2} \quad \text{for } 0 < x < 1.$$

Then $E(Y|X)$ is the random variable $X/2$. But $E(X) = 1/2$. so

$$\begin{aligned} E(Y) &= E(E(Y|X)) \\ &= E(X/2) \\ &= E(X)/2 \\ &= \frac{1}{4}. \end{aligned}$$

The first time we found this result, it involved evaluating the difficult integral $\int_0^1 y \log_e(y) dy$, the second time we found it from evaluting a double integral. Continuing, the variance of a Uniform $(0,x)$ distribution is $x^2/12$. We have

$$\begin{aligned} \mathrm{Var}(Y) &= E(\mathrm{Var}(Y|X)) + \mathrm{Var}(E(Y|X)) \\ &= E(X^2/12) + \mathrm{Var}(X/2) \\ &= E(X^2)/12 + \mathrm{Var}(X)/4 \\ &= \frac{1}{3 \cdot 12} + \frac{1}{12 \cdot 4} = \frac{7}{144}. \end{aligned}$$

3.11 Application: Binomial Distribution.

In Chapter 1 we encountered the Binomial (n, θ) Distribution, the distribution of the discrete univariate random variable X,

$$\begin{aligned} b(x; n, \theta) &= \binom{n}{x} \theta^x (1-\theta)^{n-x} \quad \text{for } x = 0, 1, 2, \ldots, n \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

It is possible to evaluate $E(X)$ and $\mathrm{Var}(X)$ directly from the definitions, with some cleverness. Paradoxically, it is much easier to evaluate them by considering X as a transformation of a multivariate random variable. The random variable X is the total number of successes in n independent trials. Let X_i be the number of successes on the i^{th} trial; that is,

$$\begin{aligned} X_i &= 1 \quad \text{if } i^{th} \text{ trial is a success} \\ &= 0 \quad \text{if } i^{th} \text{ trial is a failure.} \end{aligned}$$

Then $(X_1, X_2, \ldots, X_n)$ is a multivariate random variable of a particularly simple kind, and all the X_i's are mutually independent. Furthermore,

$$X = \sum_{i=1}^{n} X_i.$$

Now

$$E(X_i) = 1 \cdot \theta + 0 \cdot (1-\theta) = \theta,$$

$$E(X_i^2) = 1^2 \cdot \theta + 0^2 \cdot (1-\theta) = \theta,$$

and

$$\mathrm{Var}(X_i) = \theta - \theta^2 = \theta(1-\theta).$$

From (3.48) and (3.51) we have

$$\begin{aligned} E(X) &= \sum_{i=1}^{n} E(X_i) \qquad (3.67) \\ &= \sum_{i=1}^{n} \theta = n\theta, \\ \mathrm{Var}(X) &= \sum_{i=1}^{n} \mathrm{Var}(X_i) \qquad (3.68) \\ &= n\theta(1-\theta). \end{aligned}$$

With these basic formulae, let us take another look at Bayes's Billiard Table (Example 3.D). There we had Y as a continuous random variable with a Uniform $(0, 1)$ distribution, and, given $Y = y$, X had a Binomial (n, y) distribution. We found that the marginal distribution of X was a discrete uniform distribution,

$$\begin{aligned} p_X(x) &= \frac{1}{n+1} \quad \text{for } 0, 1, \ldots, n \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

We have just established that

$$E(X|Y = y) = ny,$$

$$\mathrm{Var}(X|Y = y) = ny(1-y).$$

Thus $E(X|Y)$ and $\mathrm{Var}(X|Y)$ are the continuous random variables nY and $nY(1-Y)$ respectively. From (3.65) and (3.66) we have

$$\begin{aligned} E(X) &= E[E(X|Y)] \\ &= E(nY) \\ &= n \cdot E(Y) \\ &= \frac{n}{2}, \\ \mathrm{Var}(X) &= E(\mathrm{Var}(X|Y)) + \mathrm{Var}(E(X|Y)) \\ &= E(nY(1-Y)) + \mathrm{Var}(nY) \\ &= nE(Y) - nE(Y^2) + n^2 \,\mathrm{Var}(Y) \\ &= \frac{n}{2} - \frac{n}{3} + \frac{n^2}{12} = \frac{n(n+2)}{12}. \end{aligned}$$

These agree with the results found in Example 2.0.

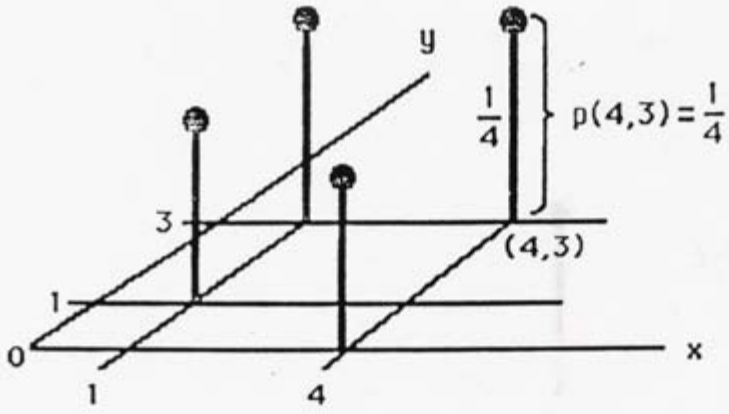


Figure 3.1. A discrete bivariate distribution.

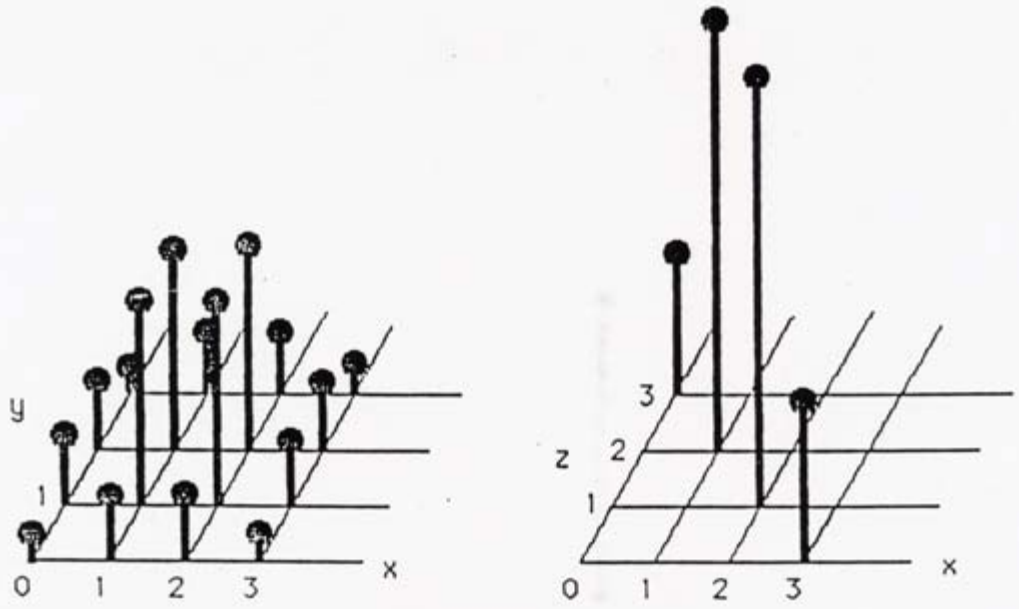
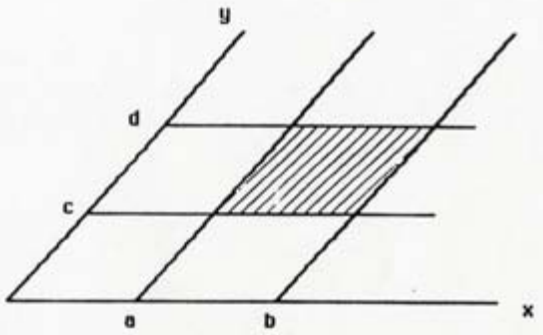


Figure 3.2. Two bivariate distributions with the same binomial marginal distributions.



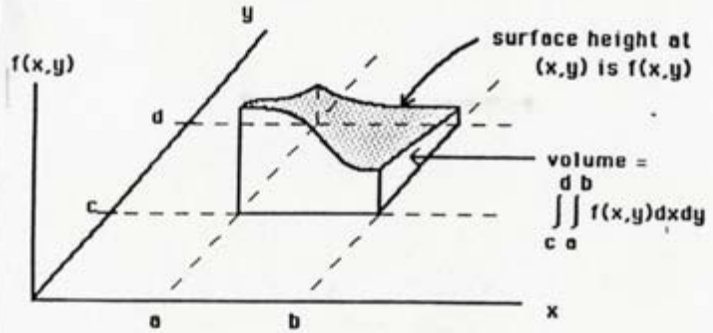
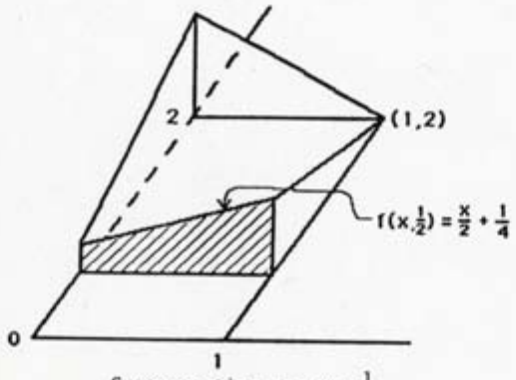
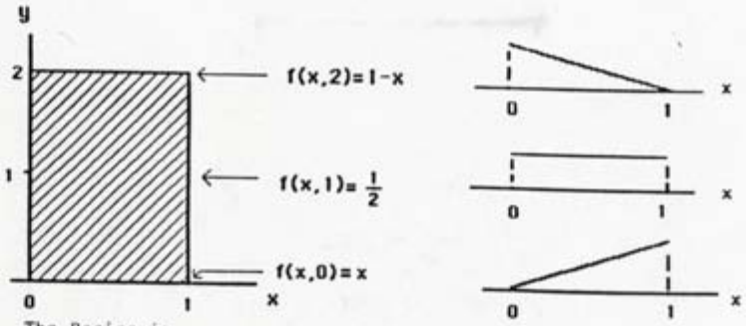


Figure 3.3. A bivariate density.



Cross-section at $y = \frac{1}{2}$.

Cross-sections



The Region in the x - y plane for which $f(x, y) > 0$.

Figure 3.4. A bivariate density and its cross-sections.

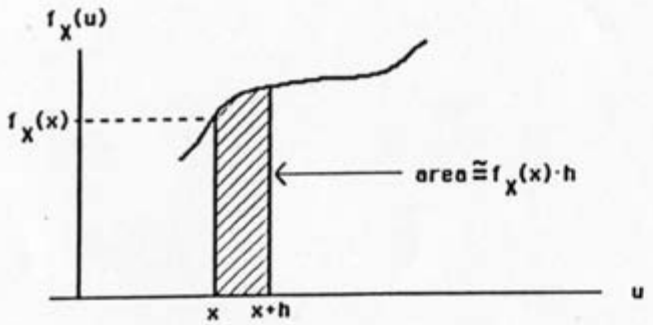
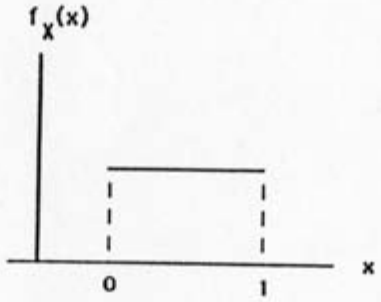
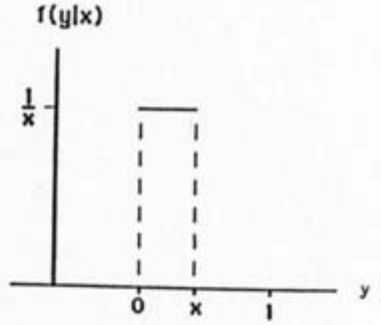
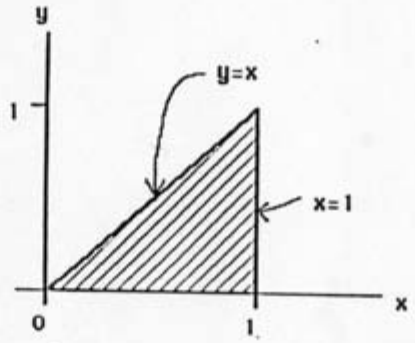


Figure 3.5







Region where $f(x, y) > 0$

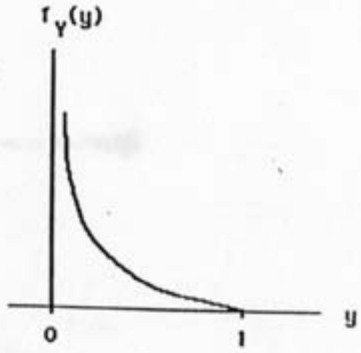


Figure 3.6. The number guessing example.

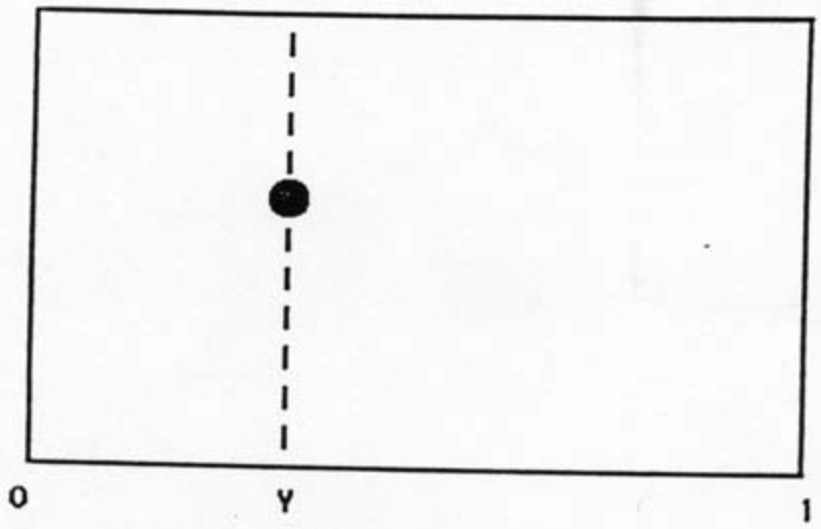


Figure 3.7. Bayes's billiard table.

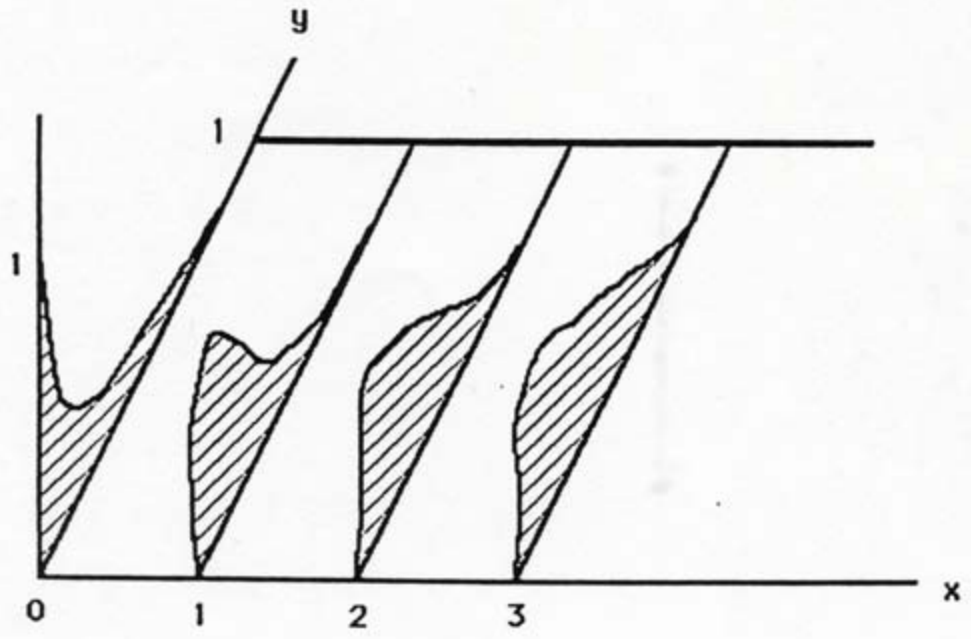
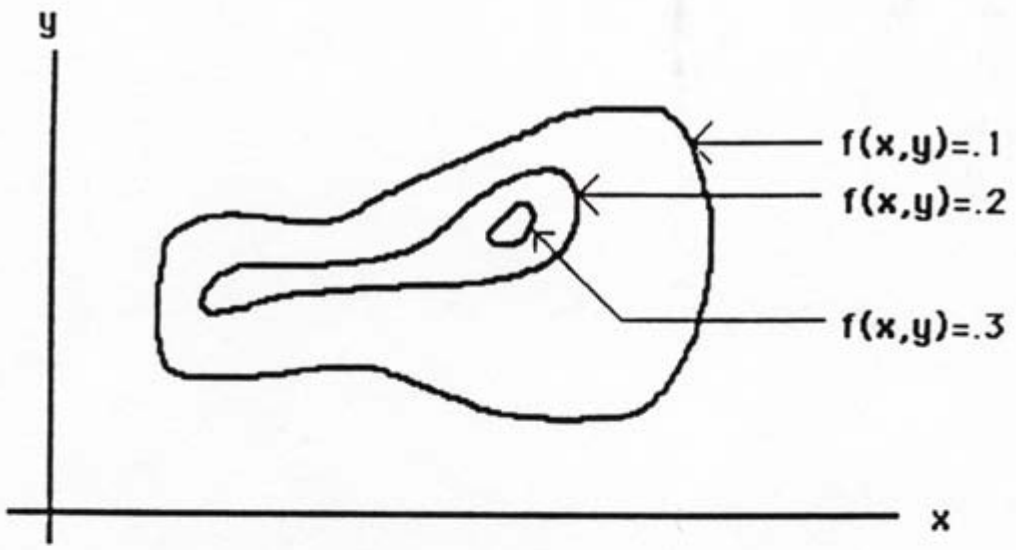


Figure 3.8



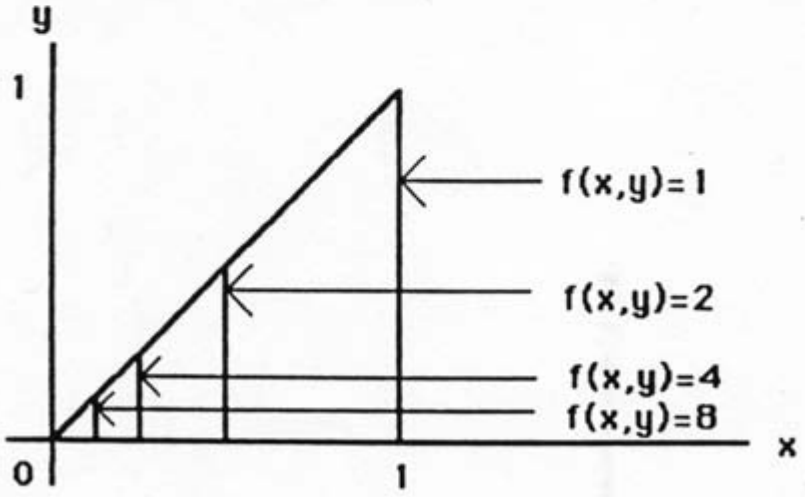


Figure 3.9

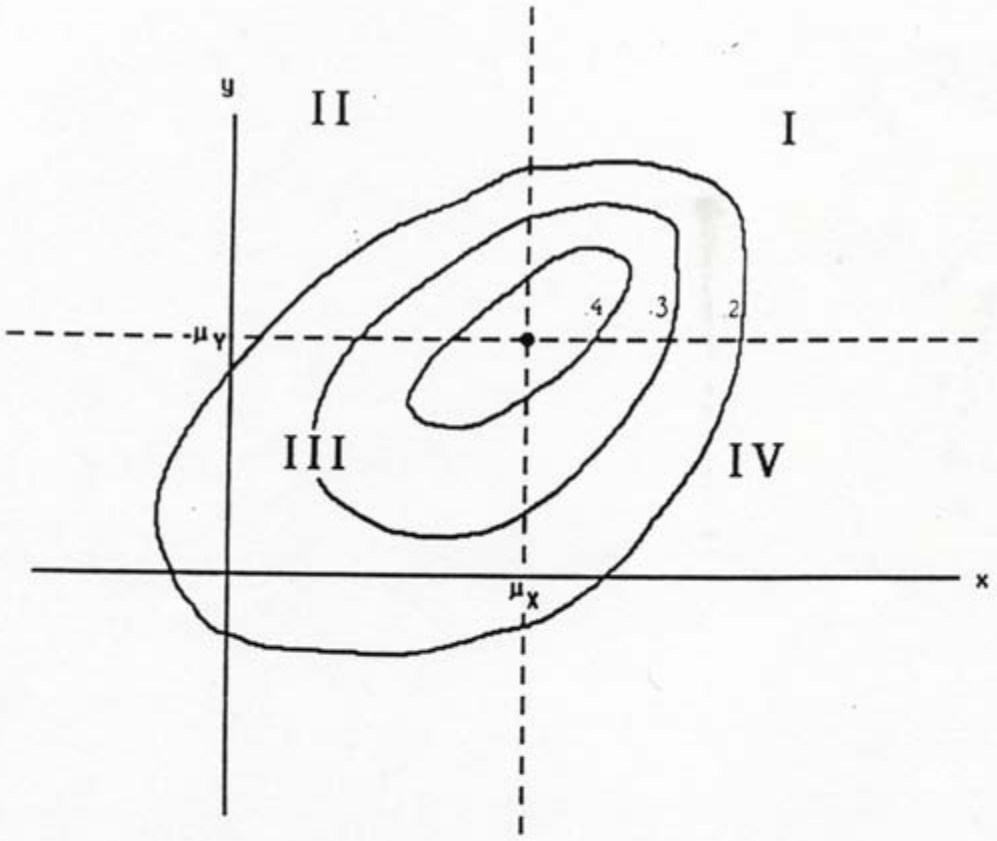


Figure 3.10

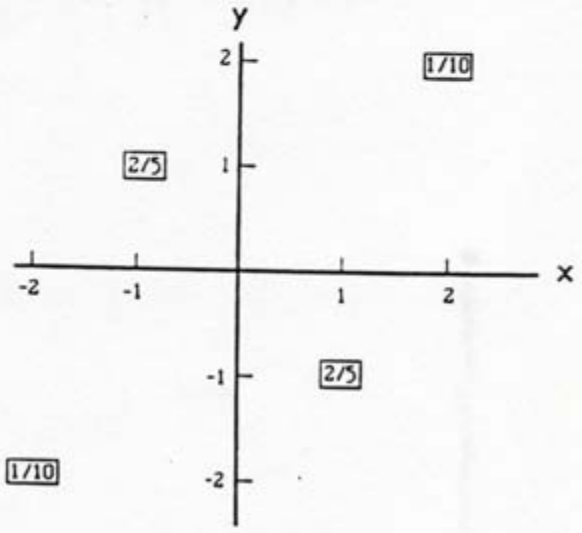


Figure 3.11. An example with Cov(X, Y) = 0 but a visual appearance of linear association.