Queen Mary
University of London

MAS 108 Probability I

Notes 11 Autumn 2005

Two discrete random variables

If X and Y are discrete random variables defined on the same sample space, then events such as

$$\text{“}X = x \text{ and } Y = y\text{”}$$

are well defined. The *joint distribution* of X and Y is a list of the probabilities

$$p_{X,Y}(x,y) = P(X = x \text{ and } Y = y)$$

for all values of x and y that occur. The list is usually shown as a two-way table. This table is called the *joint probability mass function* of X and Y. We abbreviate $p_{X,Y}(x,y)$ to $p(x,y)$ if X and Y can be understood from the context.

Example (Muffins: part 1) I have a bag containing 5 chocolate-chip muffins, 3 blueberry muffins and 2 lemon muffins. I choose three muffins from this bag at random. Let X be the number of lemon muffins chosen and Y be the number of blueberry muffins chosen.

We calculate the values of the joint probablity mass function. First note that $|\mathcal{S}| = {}^{10}\mathrm{C}_3 = 120$. Then

$$P(X = 0 \text{ and } Y = 0) = P(3 \text{ chocolate-chip are chosen}) = \frac{{}^5\mathrm{C}_3}{120} = \frac{10}{120}.$$

$$\begin{aligned} P(X = 0 \text{ and } Y = 1) &= P(1 \text{ blueberry and } 2 \text{ chocolate chip are chosen}) \\ &= \frac{{}^3\mathrm{C}_1 \times {}^5\mathrm{C}_2}{120} = \frac{30}{120}. \end{aligned}$$

The other values are found by similar calculations, giving the following table.

		Values of Y			
		0	1	2	3
Values of X	0	$\frac{10}{120}$	$\frac{30}{120}$	$\frac{15}{120}$	$\frac{1}{120}$
	1	$\frac{20}{120}$	$\frac{30}{120}$	$\frac{6}{120}$	0
	2	$\frac{5}{120}$	$\frac{3}{120}$	0	0

Of course, we check that the probabilities in the table add up to 1.

To obtain the distributions of X and Y individually, find the row sums and the column sums: these give their probability distributions. The *marginal probability mass function* p_X of X is the list of probabilities

$$p_X(x) = P(X = x) = \sum_y p_{X,Y}(x,y)$$

for all values of x which occur. Similarly, the marginal probability mass function p_Y of Y is given by

$$p_Y(y) = P(Y = y) = \sum_x p_{X,Y}(x,y).$$

The distributions of X and Y are said to be *marginal* to the joint distribution. They are just ordinary distributions.

Example (Muffins: part 2) Here the marginal p.m.f. of X is

x	0	1	2
$p_X(x)$	$\frac{7}{15}$	$\frac{7}{15}$	$\frac{1}{15}$

and the marginl p.m.f. of Y is

y	0	1	2	3
$p_Y(y)$	$\frac{35}{120}$	$\frac{63}{120}$	$\frac{21}{120}$	$\frac{1}{120}$

.

Therefore $E(X) = 3/5$ and $E(Y) = 9/10$.

We can define events in terms of X and Y, such as "$X < Y$" or "$X + Y = 3$". To find the probability of such an event, find the probability of the set of all pairs (x,y) for which the statement is true.

We can also define new random variables as functions of X and Y.

Proposition 10 If g is a real function of two variables then

$$E(g(X,Y)) = \sum_x \sum_y g(x,y) p_{X,Y}(x,y).$$

The proof is just like the proof of Proposition 8.

Example (Muffins: part 3) Put $U = X + Y$ and $V = XY$. Then

$$\begin{aligned} E(U) &= 1 \times \frac{30}{120} + 2 \times \frac{15}{120} + 3 \times \frac{1}{120} + 1 \times \frac{20}{120} + 2 \times \frac{30}{120} + 3 \times \frac{6}{120} + 2 \times \frac{5}{120} + 3 \times \frac{3}{120} \\ &= \frac{3}{2} = E(X) + E(Y). \end{aligned}$$

However,

$$E(V) = 1 \times \frac{30}{120} + 2 \times \frac{6}{120} + 2 \times \frac{3}{120} = \frac{2}{5} \neq E(X)E(Y).$$

Theorem 7 (a) $E(X+Y) = E(X) + E(Y)$ always.

(b) $E(XY)$ is not necessarily equal to $E(X)E(Y)$.

Proof (a)

$$\begin{aligned} E(X+Y) &= \sum_x \sum_y (x+y) p_{X,Y}(x,y) \qquad \text{by Proposition 10} \\ &= \sum_x \sum_y x p_{X,Y}(x,y) + \sum_x \sum_y y p_{X,Y}(x,y) \\ &= \sum_x x \sum_y p_{X,Y}(x,y) + \sum_y y \sum_x p_{X,Y}(x,y) \\ &= \sum_x x p_X(x) + \sum_y y p_Y(y) \\ &= E(X) + E(Y). \end{aligned}$$

(b) A single counter-example is all that we need. In the muffins example we saw that $E(XY) \neq E(X)E(Y)$. ■

Independence

Random variables X and Y defined on the same sample space are defined to be *independent* of each other if the events "$X = x$" and "$Y = y$" are independent for all values of x and y; that is

$$p_{X,Y}(x,y) = p_X(x) p_Y(y) \qquad \text{for all } x \text{ and } y.$$

Example (Muffins: part 4) Here $P(X=0)=7/15$ and $P(Y=0)=35/120$ but

$$P(X=0 \text{ and } Y=0)=\frac{10}{120}\neq\frac{7}{15}\times\frac{35}{120}=P(X=0)P(Y=0),$$

so X and Y are not independent of each other. Note that a single pair of values of x and y where the probabilities do not multiply is enough to show that X and Y are not independent.

On the other hand, if I roll a die twice, and X and Y are the numbers that come up on the first and second throws, then X and Y will be independent, even if the die is not fair (so that the outcomes are not all equally likely).

If we have more than two random variables (for example X, Y, Z), we say that they are *mutually independent* if the events that the random variables take specific values (for example, $X=a$, $Y=b$, $Z=c$) are mutually independent.

Theorem 8 If X and Y are independent of each other then $E(XY)=E(X)E(Y)$.

Proof

$$\begin{aligned}
E(XY) &= \sum_x\sum_y xyp_{X,Y}(x,y)\\
&= \sum_x\sum_y xyp_X(x)p_Y(y) \qquad \text{if } X \text{ and } Y \text{ are independent}\\
&= \left[\sum_x xp_X(x)\right]\left[\sum_y yp_Y(y)\right]\\
&= E(X)E(Y). \quad \blacksquare
\end{aligned}$$

Note that the converse is not true, as the following example shows.

Example Suppose that X and Y have the joint p.m.f. in this table.

		Values of Y		
		-1	0	1
Values of X	0	0	$\frac{1}{2}$	0
	1	$\frac{1}{4}$	0	$\frac{1}{4}$

The marginal distributions are

x	0	1
$p_X(x)$	$\frac{1}{2}$	$\frac{1}{2}$

and

y	-1	0	1
$p_Y(y)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Therefore $P(X=0 \text{ and } Y=1)=0$ but $P(X=0)P(Y=1)=\frac{1}{2}\times\frac{1}{4}\neq 0$ so X and Y are not independent of each other. However, $E(X)=1/2$ and $E(Y)=0$ so $E(X)E(Y)=0$, while

$$E(XY)=-1\times\frac{1}{4}+0\times\frac{1}{2}+1\times\frac{1}{4}=0,$$

so $E(XY)=E(X)E(Y)$.

Covariance and correlation

A measure of the joint spread of X and Y is the *covariance* of X and Y, which is defined by

$$\text{Cov}(X,Y)=E\left((X-\mu_X)(Y-\mu_Y)\right),$$

where $\mu_X=E(X)$ and $\mu_Y=E(Y)$.

Theorem 9 (Properties of covariance) (a) $\text{Cov}(X,X)=\text{Var}(X)$.

(b) $\text{Cov}(X,Y)=E(XY)-E(X)E(Y)$.

(c) If a is a constant then $\text{Cov}(aX,Y)=\text{Cov}(X,aY)=a\,\text{Cov}(X,Y)$.

(d) If b is constant then $\text{Cov}(X,Y+b)=\text{Cov}(X+b,Y)=\text{Cov}(X,Y)$.

(e) If X and Y are independent, then $\text{Cov}(X,Y)=0$.

Proof (a) This follows directly from the definition.

(b)

$$\begin{aligned}
\text{Cov}(X,Y) &= E\left((X-\mu_X)(Y-\mu_Y)\right)\\
&= E\left(XY-\mu_X Y-\mu_Y X+\mu_X\mu_Y\right)\\
&= E(XY)+E(-\mu_X Y)+E(-\mu_Y X)+E(\mu_X\mu_Y) \qquad \text{by Theorem 7(a)}\\
&= E(XY)-\mu_X E(Y)-\mu_Y E(X)+\mu_X\mu_Y \qquad \text{by Theorem 4}\\
&= E(XY)-E(X)E(Y)-E(Y)E(X)+E(X)E(Y)\\
&= E(XY)-E(X)E(Y).
\end{aligned}$$

(c) From part (b), $\text{Cov}(aX,Y) = E(aXY) - E(aX)E(Y)$. From Theorem 4, $E(aXY) = aE(XY)$ and $E(aX) = aE(X)$. Therefore

$$\text{Cov}(aX,Y) = aE(XY) - aE(X)E(Y) = a(E(XY) - E(X)E(Y)) = a\,\text{Cov}(X,Y).$$

(d) Theorem 4 shows that $E(Y+b) = E(Y)+b$, so we have $(Y+b) - E(Y+b) = Y - E(Y)$. Therefore

$$\begin{aligned} \text{Cov}(X,Y+b) &= E(X-E(X))(Y+b-E(Y+b)) \\ &= E(X-E(X))(Y-E(Y)) = \text{Cov}(X,Y). \end{aligned}$$

(e) From part (b), $\text{Cov}(X,Y) = E(XY) - E(X)E(Y)$. Theorem 8 shows that $E(XY) - E(X)E(Y) = 0$ if X and Y are independent of each other. ■

The covariance of X and Y is the product of the distance of X from its mean and the distance of Y from its mean. Therefore, if $X - \mu_X$ and $Y - \mu_Y$ tend to be either both postive or both negative then $\text{Cov}(X,Y)$ is positive. On the other hand, if $X - \mu_X$ and $Y - \mu_Y$ tend to have opposite signs then $\text{Cov}(X,Y)$ is negative.

If $\text{Var}(X)$ or $\text{Var}(Y)$ is large then $\text{Cov}(X,Y)$ may also be large; in fact, multiplying X by a constant a multiplies $\text{Var}(X)$ by a^2 and $\text{Cov}(X,Y)$ by a. To avoid this dependence on scale, we define the *correlation coefficient*, $\text{corr}(X,Y)$, which is just a "normalised" version of the covariance. It is defined as follows:

$$\text{corr}(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}}.$$

The point of this is the first and last parts of the following theorem.

Theorem (Properties of correlation) Let X and Y be random variables. Then

(a) $-1 \le \text{corr}(X,Y) \le 1$;

(b) if X and Y are independent, then $\text{corr}(X,Y) = 0$;

(c) if $Y = mX + c$ for some constants $m \neq 0$ and c, then $\text{corr}(X,Y) = 1$ if $m > 0$, and $\text{corr}(X,Y) = -1$ if $m < 0$; this is the only way in which $\text{corr}(X,Y)$ can be equal to ± 1;

(d) $\text{corr}(X,Y)$ is independent of the units of measurement in the sense that if X or Y is multilpied by a constant, or has a constant added to it, then $\text{corr}(X,Y)$ is unchanged.

The proof of the first part won't be given here. But note that this is another check on your calculations: if you calculate a correlation coefficient which is bigger than 1 or smaller than -1, then you have made a mistake. Part (b) follows immediately from part (e) of the preceding theorem.

For part (c), suppose that $Y = mX + c$. Let $\mathrm{Var}(X) = \alpha$. Now we just calculate everything in sight.

$$\mathrm{Var}(Y) = \mathrm{Var}(mX + c) = \mathrm{Var}(mX) = m^2 \mathrm{Var}(X) = m^2\alpha$$

$$\mathrm{Cov}(X,Y) = \mathrm{Cov}(X, mX + c) = \mathrm{Cov}(X, mX) = m\,\mathrm{Cov}(X,X) = m\alpha$$

$$\begin{aligned} \mathrm{corr}(X,Y) &= \mathrm{Cov}(X,Y)/\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)} \\ &= m\alpha/\sqrt{m^2\alpha^2} \\ &= \begin{cases} +1 & \text{if } m > 0, \\ -1 & \text{if } m < 0. \end{cases} \end{aligned}$$

The proof of the converse will not be given here.

Part (d) follows from Theorems 4, 5 and 9.

Thus the correlation coefficient is a measure of the extent to which the two variables are linearly related. It is $+1$ if Y increases linearly with X; 0 if there is no linear relation between them; and -1 if Y decreases linearly as X increases. More generally, a positive correlation indicates a tendency for larger X values to be associated with larger Y values; a negative value, for smaller X values to be associated with larger Y values.

We call two random variables X and Y *uncorrelated* if $\mathrm{Cov}(X,Y) = 0$ (in other words, if $\mathrm{corr}(X,Y) = 0$). The preceding theorem and example show that we can say:

> Independent random variables are uncorrelated, but uncorrelated random variables need not be independent.

Example In some years there are two tests in the Probability class. We can take as the sample space the set of all students who take both tests, and choose a student at random. Put $X(s) =$ the mark of student s on Test 1 and $Y(s) =$ the mark of student s on Test 2. We would expect a student who scores better than average on Test 1 to do so again on Test 2, and one who scores worse than average to do so again on Test 2, so there should be a positive correlation between X and Y. However, we would not expect the Test 2 scores to be perfectly predictable as a linear function of the Test 1 scores. The marks for one particular year are shown in Figure 1. The correlation is 0.69 to 2 decimal places.

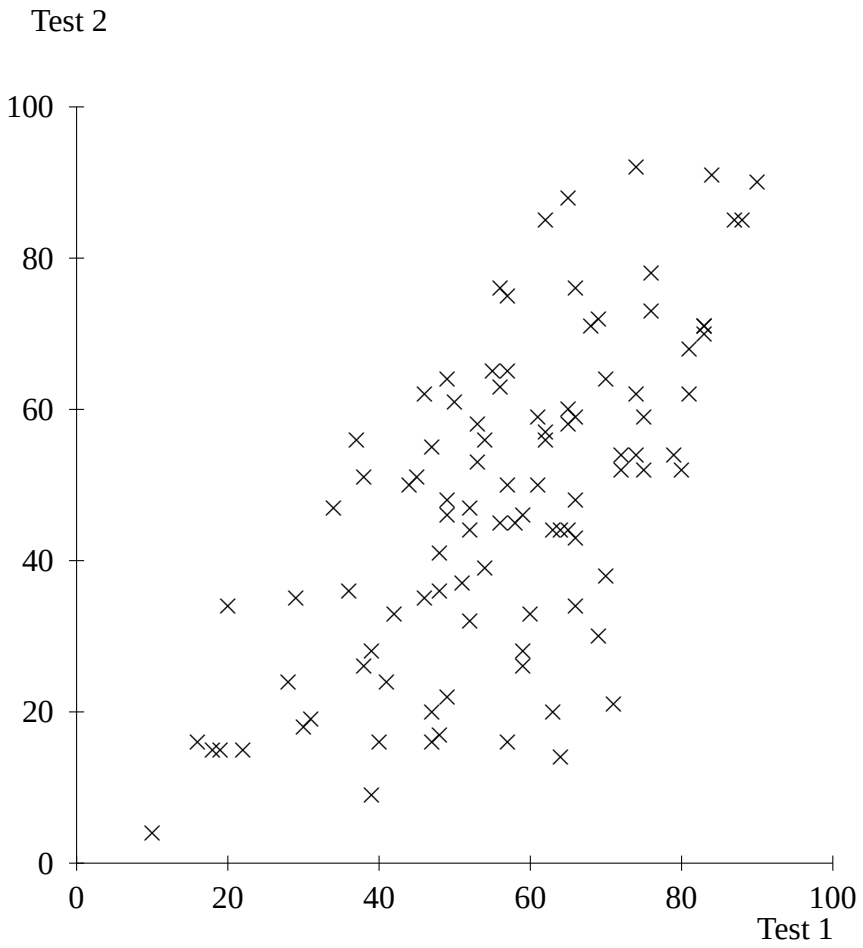

Figure 1: Marks on two Probability tests

Theorem 10 If X and Y are random variables and a and b are constants then

$$\mathrm{Var}(aX+bY) = a^2\,\mathrm{Var}(X) + 2ab\,\mathrm{Cov}(X,Y) + b^2\,\mathrm{Var}(Y).$$

Proof First,

$$\begin{aligned} E(aX+bY) &= E(aX)+E(bY) && \text{by Theorem 7} \\ &= aE(X)+bE(Y) && \text{by Theorem 4} \\ &= a\mu_X + b\mu_Y. \end{aligned}$$

Then

$$\begin{aligned} \mathrm{Var}(aX+bY) &= E[(aX+bY)-E(aX+bY)]^2 && \text{by definition of variance} \\ &= E[aX+bY-a\mu_X-b\mu_Y]^2 \\ &= E[a(X-\mu_X)+b(Y-\mu_Y)]^2 \\ &= E[a^2(X-\mu_X)^2+2ab(X-\mu_X)(Y-\mu_Y)+b^2(Y-\mu_Y)^2] \\ &= E[a^2(X-\mu_X)^2]+E[2ab(X-\mu_X)(Y-\mu_Y)]+E[b^2(Y-\mu_Y)^2] \\ &\qquad \text{by Theorem 7} \\ &= a^2E[(X-\mu_X)^2]+2abE[(X-\mu_X)(Y-\mu_Y)]+b^2E[(Y-\mu_Y)^2] \\ &\qquad \text{by Theorem 4} \\ &= a^2\,\mathrm{Var}(X)+2ab\,\mathrm{Cov}(X,Y)+b^2\,\mathrm{Var}(Y). \quad \blacksquare \end{aligned}$$

Corollary If X and Y are independent, then

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

(b) $\mathrm{Var}(X-Y) = \mathrm{Var}(X)+\mathrm{Var}(Y)$.

Mean and variance of binomial

Here is the third way of finding the mean and variance of a binomial distribution, which is more sophisticated than the methods in Notes 8 yet easier than both. If $X \sim \text{Bin}(n, p)$ then X can be thought of as the the number of successes in n mutually independent Bernoulli trials, each of which has probability p of success.

For $i = 1, \ldots, n$, let X_i be the number of successes at the i-th trial. Then $X_i \sim \text{Bernoulli}(p)$, so $E(X_i) = p$ and $\text{Var}(X_i) = pq$, where $q = 1 - p$. Moreover, the random variables $X_1, \ldots, X_n$ are mutually independent.

By Theorem 7,

$$\begin{aligned} E(X) &= E(X_1) + E(X_2) + \cdots + E(X_n) \\ &= \underbrace{p + p + \cdots + p}_{n \text{ times}} \\ &= np. \end{aligned}$$

By the Corollary to Theorem 10,

$$\begin{aligned} \text{Var}(X) &= \text{Var}(X_1) + \text{Var}(X_2) + \cdots + \text{Var}(X_n) \\ &= \underbrace{pq + pq + \cdots + pq}_{n \text{ times}} \\ &= npq. \end{aligned}$$

Note that if also $Y \sim \text{Bin}(m, p)$, and X and Y are independent of each other, then $Y = X_{n+1} + \cdots + X_{n+m}$, where $X_{n+1}, \ldots, X_{n+m}$ are mutually independent Bernoulli(p) random variables which are all independent of $X_1, \ldots, X_n$, so

$$X + Y = X_1 \cdots + X_n + X_{n+1} + \cdots X_{n+m},$$

which is the sum of $n + m$ mutually independent Bernoulli(p) random variables, so $X + Y \sim \text{Bin}(n + m, p)$.