Jointly distributed random variables

Tan Do

Vietnamese-German University

Lecture 10

In this lecture

- Joint probability distributions and marginal probability distributions
- Independent random variables
- Covariance of random variables

Joint probability distributions

Sometimes it is desired to consider 2 random variables X and Y at the same time. For that, we can use their joint probability distribution.

- If X and Y are discrete:

  Their joint probability mass function consists of probability values $P(X = x_i, Y = y_j) = p_{ij} \geq 0$ satisfying

$$\sum_i \sum_j p_{ij} = 1.$$

  The joint cumulative distribution function is defined to be

$$F(x, y) = P(X \leq x, Y \leq y) = \sum_{i: x_i \leq x} \sum_{j: y_j \leq y} p_{ij}.$$

Example (Air conditioner maintenance) A company that services air conditioners is interested in how to schedule its technicians in the most efficient manner. Specifically, the company is interested in how long a technician takes on a visit to a particular location. The company recognizes that this depends on the number of air conditioners at the location that needs to be serviced.

Let the random variable X (taking values 1, 2, 3 and 4) be the *service time* in hours. Let the random variable Y (taking values 1, 2 and 3) be the *number of air conditioners* at a location. Then X and Y are jointly distributed. Suppose their joint probability mass function is given by the tables.

Y = number of air conditioner units \ X = service time (hrs)	1	2	3	4
1	0.12	0.08	0.07	0.05
2	0.08	0.15	0.21	0.13
3	0.01	0.01	0.02	0.07

$$\sum_i \sum_j p_{ij} = 0.12 + 0.08 + \cdots + 0.07 = 1$$

The joint cumulative distribution function

$$F(x,y) = P(X \leq x, Y \leq y) = \sum_{i=1}^{x} \sum_{j=1}^{y} p_{ij}$$

is given by the next figure.

Y = number of air conditioner units \ X = service time (hrs)	1	2	3	4
1	0.12	0.20	0.27	0.32
2	0.20	0.43	0.71	0.89
3	0.21	0.45	0.75	1.00

The joint cumulative distribution function

For example, the probability that a location has no more than 2 air conditioners that take no more than 2 hours to service is $F(2,2) = p_{11} + p_{12} + p_{21} + p_{22} = 0.43$.

Marginal probability distributions

Suppose 2 randoms variables X and Y are jointly distributed. Suppose we want to calculate the probability $P(X = x_i)$ (regardless of the value of Y).

- If X and Y are discrete:

$$P(X = x_i) = p_{i+} = \sum_j p_{ij}.$$

  This gives the marginal probability mass function of X.

The marginal p.m.f of Y are defined similarly.

The expectations and variances of the random variables X and Y can be obtained from their marginal distributions in the usual manner.

Example (Air conditioner maintenance) Recall that X is the time taken to service the air conditioners at a particular location.

We have

$$P(X=1) = \sum_{j=1}^{3} p_{1j} = 0.12 + 0.08 + 0.01 = 0.21.$$

The expected service time is

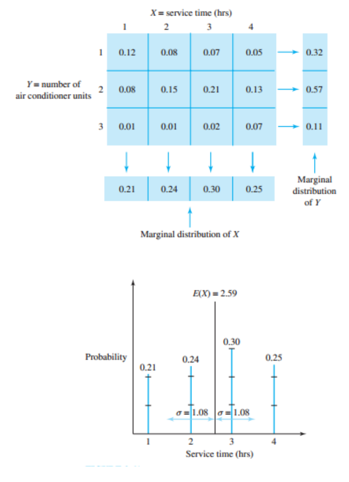
$$E(X) = \sum_{i=1}^{4} i\, P(X=i)$$

$$= 1 \times 0.21 + 2 \times 0.24 + 3 \times 0.3 + 4 \times 0.25 = 2.59.$$

Also, $E(X^2) = \sum_{i=1}^{4} i^2\, P(X=i) = 7.87$. So

$$\mathrm{Var}(X) = E(X^2) - E(X)^2 = 7.87 - 2.59^2 = 1.162.$$

This gives $\sigma_X = \sqrt{\mathrm{Var}(X)} = 1.08$ hours.



Exercise Calculate the marginal p.m.f of Y, $E(Y)$ and $\mathrm{Var}(Y)$.

Conditional probability distributions

It is sometimes useful to consider the distribution of one random variable *conditional* on another random variable.

Let X and Y be 2 random variables which are jointly distributed.

- If X and Y are discrete: The conditional distribution (p.m.f) of X conditional on the event $Y = y_j$ consists of the probability values

$$p_{i|Y=y_j} = P(X = x_i|Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} = \frac{p_{ij}}{p_{+j}},$$

where $p_{+j} = P(Y = y_j) = \sum_i p_{ij}$.

Example (Air conditioner maintenance) Let Y be the event that a technician is visiting a location that is known to have 3 air conditioners. Then

$$P(Y=3)=p_{+3}=p_{13}+p_{23}+p_{33}+p_{43}=0.11.$$

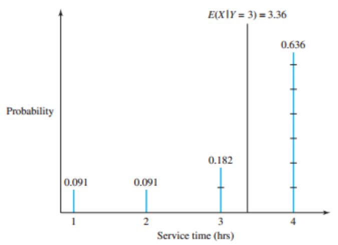
The conditional distribution of the service time X consists of the probability values

$$p_{1|Y=3}=P(X=1|Y=3)=\frac{p_{13}}{p_{+3}}=\frac{0.01}{0.11}=0.091,$$

$$p_{2|Y=3}=P(X=2|Y=3)=\frac{p_{23}}{p_{+3}}=\frac{0.01}{0.11}=0.091,$$

$$p_{3|Y=3}=P(X=3|Y=3)=\frac{p_{33}}{p_{+3}}=\frac{0.02}{0.11}=0.182,$$

$$p_{4|Y=3}=P(X=4|Y=3)=\frac{p_{43}}{p_{+3}}=\frac{0.07}{0.11}=0.636.$$



The results imply that conditioning on a location having 3 air conditioners increases the chances of a large service time being required.

The conditional expectation of the service time is

$$E(X|Y=3) = \sum_{i=1}^{4} i p_{i|Y=3} = 1 \times 0.091 + 2 \times 0.091 + 3 \times 0.182 + 4 \times 0.636 = 3.36.$$

Note that we calculated before that

$$E(X) = 2.59.$$

So there is a significant difference in the "overall" expected service time.

Reason If a technician sets off for a location whose number of air conditioners is not known, then the expected service time at the location is 2.59 hours.

However, if the technician knows that there are 3 air conditioners at the location that need servicing, then the expected service time is 3.36 hours.

Independence

Next we consider the independent property of two random variables X and Y.

- If X and Y are discrete: X and Y are independent if their joint p.m.f is the *product* of their 2 marginal distribution, i.e.,

$$p_{ij} = p_{i+} \times p_{+j}$$

  for all values x_i and y_j.

If 2 random variables are independent, then the probability distribution of one random variable does not depend on the other and vice versa.

Note that this definition is similar to the independent property of 2 events.

A consequence of the independence of 2 random variables is that their *conditional* distributions are identical to their *marginal* distributions:

- If X and Y are discrete, then

$$p_{i|Y=y_j} = P(X = x_i|Y = y_j) = \frac{p_{ij}}{p_{+j}} = \frac{p_{i+}p_{+j}}{p_{+j}} = p_{i+}.$$

Similarly $p_{j|X=x_i} = P(Y = y_j|X = x_i) = p_{+j}$.

Games of chance

Coin tossing Suppose that a fair coin is tossed 3 times so that there are 8 equally likely outcomes.

Let X be the random variable that gives the number of heads in the *first* and *second* tosses.

Let Y be the random variable that gives the number of heads in the *third* toss.

The joint probability of X and Y is given by the figure.

For example,

$$P(X=0,Y=0)=P(TTT)=\frac{1}{8}$$

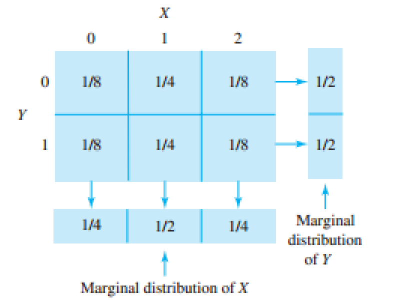
and

$$P(X=0)=P(TTT)+P(TTH)=\frac{1}{4}.$$

Check to see that

$P(X=i,Y=j)=P(X=i)P(Y=j)$

so that X and Y are independent.



Let Z be the random variable that gives the number of head in the *second* and *third* tosses.

The joint probability of X and Z is given by the figure.

For example,

$$P(X=1, Z=1) = P(HTH) + P(THT) = \frac{1}{4}.$$

But notice that

$$P(X=0, Z=0) = P(TTT) = \frac{1}{8},$$

$$P(X=0) = P(TTH) + P(TTT) = \frac{1}{4}$$

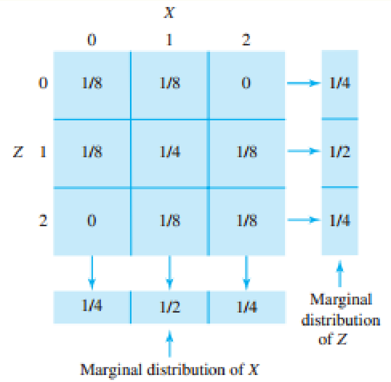
and

$$P(Z=0) = P(HTT) + P(TTT) = \frac{1}{4}.$$

Therefore

$P(X=0, Z=0) \neq P(X=0)P(Z=0)$

so that X and Z are *not* independent.



Covariance

The covariance of 2 random variables X and Y is defined by

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

The covariance can be either a positive or a negative number.

Special case   If X and Y are independent, then $\mathrm{Cov}(X,Y) = 0$.

So $\mathrm{Cov}(X,Y)$ can be used to indicate the independence of X and Y.

In practice, the most convenient way to assess the strength of the dependence between 2 random variables is through their correlation.

The correlation between 2 random variables X and Y is defined by

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

The correlation takes values between -1 and 1.

Note that if X and Y are independent, $\mathrm{Corr}(X,Y) = 0$.

Random variables with a positive correlation are said to be positively correlated. In such cases, there is a tendency for *high* values of one random variable to be associated with *high* values of the other random variable.

Random variables with a negative correlation are said to be negatively correlated. In such cases, there is a tendency for *high* values of one random variable to be associated with *low* values of the other random variable.

The strength of these tendencies increases as the correlation moves further away from 0 to 1 or to -1.

Example (Air conditioner maintenance) The expected service time is $E(X) = 1.79$.

The expected number of air conditioners serviced is $E(Y) = 1.79$.

In addition,

$$E(XY) = \sum_{i=1}^{4} \sum_{j=1}^{3} i\, j\, p_{ij} = 1 \times 1 \times 0.12 + 1 \times 2 \times 0.08 + \ldots + 4 \times 3 \times 0.07 = 4.86.$$

The covariance is

$$\mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y) = 4.86 - 2.59 \times 1.79 = 0.224.$$

Since $\mathrm{Var}(X) = 1.162$ and $\mathrm{Var}(Y) = 0.386$, the correlation between the service time and the number of air conditioners serviced is

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{0.224}{\sqrt{1.162 \times 0.386}} = 0.334.$$

Exercise Interpret the meaning of $\mathrm{Corr}(X, Y) = 0.334$ in the example.

Answer Since $\mathrm{Corr}(X, Y) = 0.334$, the service time and the number of air conditioners serviced are not independent, but are positively correlated.

This is expected since it is clear that locations with a large number of air conditioners requires long service times.

Exercise Calculate the correlations of the random variables X, Y and Z in the Coin Tossing example. Interpret your results.