5 — CORRELATION

The covariance of two random variables gives some measure of their independence. A second way of assessing the measure of independence will be discussed shortly but first the expectation and variance of the Binomial distribution will be determined. The calculations turn out to be surprisingly tedious.

Expectation of the Binomial Distribution — I

The Binomial distribution can be defined as:

$$\mathrm{P}(X = r) = \binom{n}{r} p^r q^{n-r} \qquad \text{where } p + q = 1 \text{ and } 0 \leqslant r \leqslant n$$

The expectation is:

$$\mathrm{E}(X) = \sum_{r=0}^{n} r.\mathrm{P}(X = r) = \sum_{r=0}^{n} r.\binom{n}{r} p^r q^{n-r}$$

When $r = 0$ the term is zero so it is in order to begin the sum from $r = 1$:

$$\begin{aligned}
\mathrm{E}(X) &= \sum_{r=1}^{n} r \frac{n!}{r!\,(n-r)!} p^r q^{n-r} \\
&= \sum_{r=1}^{n} \frac{n!}{(r-1)!\,(n-r)!} p^r q^{n-r} \\
&= np \sum_{r=1}^{n} \frac{(n-1)!}{(r-1)!\,(n-r)!} p^{r-1} q^{n-r}
\end{aligned}$$

Next, let $s = r - 1$ which, of course, means replacing r by $s + 1$. Noting that the sum is currently from $r = 1$ to $r = n$, the replacement will mean s running from $s = 0$ to $s = n - 1$:

$$\mathrm{E}(X) = np \sum_{s=0}^{n-1} \frac{(n-1)!}{s!\,(n-s-1)!} p^s q^{n-s-1}$$

Then, let $m = n - 1$:

$$\mathrm{E}(X) = np \sum_{s=0}^{m} \frac{m!}{s!\,(m-s)!} p^s q^{m-s}$$

Finally note that by the Binomial theorem,

$$(q + p)^m = \sum_{s=0}^{m} \frac{m!}{s!\,(m-s)!} p^s q^{m-s}$$

Given that $p + q = 1$ the summation itself is 1 and hence:

$$\mathrm{E}(X) = np$$

Expectation of the Binomial Distribution — II

This seems a horribly tedious way of determining the expectation. Fortunately there is a quicker way.

Consider the most trivial Binomial distribution where a random variable is distributed Binomial$(1, p)$. A tabular representation of this trivial case is:

	$r \rightarrow$	
X	0	1
$\mathrm{P}(X = r)$	q	p

The expectation is:

$$\mathrm{E}(X) = \sum_{r=0}^{1} r.\mathrm{P}(X = r) = 0.q + 1.p = p$$

Now consider n such random variables $X_1, X_2, \ldots, X_n$. Such variables are said to be Independent and Identically Distributed, commonly simply called IID. The expectation of their sum is:

$$\mathrm{E}(X_1 + X_2 + \cdots + X_n) = \mathrm{E}(X_1) + \mathrm{E}(X_2) + \cdots + \mathrm{E}(X_n) = p + p + \cdots + p = np$$

Since the variables are Identically Distributed they each have the same expectation and the sum of these expectations is simply n times the expectation of the original variable.

Taking n variables which are each distributed Binomial$(1, p)$ is exactly equivalent to having an overall distribution Binomial(n, p). For example, in a family of four children, each child is distributed Binomial$(1, p)$ and the overall distribution is Binomial$(4, p)$.

This is clearly a quicker way of determining the expectation of the Binomial distribution and an extension of this technique leads to the variance...

Variance of the Binomial Distribution

The random variable which is distributed Binomial$(1, p)$ has expectation $\mathrm{E}(X) = p$. The expectation of the square is:

$$\mathrm{E}(X^2) = \sum_{r=0}^{1} r^2.\mathrm{P}(X = r) = 0^2.q + 1^2.p = p$$

The variance can now be computed:

$$\mathrm{V}(X) = \mathrm{E}(X^2) - \big(\mathrm{E}(X)\big)^2 = p - p^2 = p(1 - p) = pq$$

Again consider n such random variables $X_1, X_2, \ldots, X_n$. Given that they are Independent as well as Identically Distributed the variance of their sum is the sum of their variances:

$$\mathrm{V}(X_1 + X_2 + \cdots + X_n) = \mathrm{V}(X_1) + \mathrm{V}(X_2) + \cdots + \mathrm{V}(X_n) = pq + pq + \cdots + pq = npq$$

Binomial Summary

A random variable X which is distributed Binomial(n, p) is such that the probability:

$$\mathrm{P}(X = r) = \binom{n}{r} p^r q^{n-r} \qquad \text{where } p + q = 1 \text{ and } 0 \leqslant r \leqslant n$$

The expectation and variance of the random variable are:

$$\mathrm{E}(X) = np \qquad \text{and} \qquad \mathrm{V}(X) = npq$$

Correlation Coefficient

The *correlation coefficient* comes from scaling the covariance and is often referred to by the letter R:

$$\text{Correlation Coefficient} = \mathrm{R} = \frac{\mathrm{W}(X, Y)}{\sqrt{\mathrm{V}(X) \,.\, \mathrm{V}(Y)}}$$

In general $-1 \leqslant \mathrm{R} \leqslant +1$.

$\mathrm{R} = -1$	means complete negative correlation.
$\mathrm{R} = \ \ 0$	means zero correlation but not necessarily independence.
$\mathrm{R} = +1$	means complete positive correlation.

Example — Two Ordinary Dice

Let X and Y be two random variables associated with the throws of two ordinary dice. In both cases the mean is $\frac{7}{2}$ and the variance is $\frac{35}{12}$ (see page 4.2). In summary:

$$\mathrm{E}(X) = \mathrm{E}(Y) = \frac{7}{2} \qquad \mathrm{V}(X) = \mathrm{V}(Y) = \frac{35}{12}$$

To determine the covariance $\mathrm{W}(X, Y)$ it is first necessary to evaluate $\mathrm{E}(XY)$:

$$\mathrm{E}(XY) = \sum_{r=0}^{6} \sum_{s=0}^{6} r.s \, \mathrm{P}(X = r, Y = S)$$

In the case of two fair dice, the probability is always $\frac{1}{36}$ when r, s are non-zero so:

$$\mathrm{E}(XY) = \sum_{r=1}^{6} \sum_{s=1}^{6} r.s \, \frac{1}{36}$$

By the lemma on page 4.8 the summations can be separated:

$$\mathrm{E}(XY) = \left(\sum_{r=1}^{6} r \right) \left(\sum_{s=1}^{6} s \right) \frac{1}{36} = \frac{21 \times 21}{36} = \frac{49}{4}$$

The covariance and correlation can now be determined:

$$\mathrm{W}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X) \,.\, \mathrm{E}(Y) = \frac{49}{4} - \frac{7}{2} \times \frac{7}{2} = 0 \quad \text{and so } \mathrm{R} = 0$$

Example — Two Linked Dice

The result for two ordinary dice should have been obvious from the outset. The dice are independent ensuring $\mathrm{E}(XY) = \mathrm{E}(X)\,.\,\mathrm{E}(Y)$. Independence implies zero covariance and hence zero correlation.

By contrast, suppose X and Y are two random variables associated with two dice which behave as two linked drums on a broken fruit machine; both dice always show the same result. Each random variable viewed alone has the same expectation and variance as before:

$$\mathrm{E}(X) = \mathrm{E}(Y) = \frac{7}{2} \qquad \mathrm{V}(X) = \mathrm{V}(Y) = \frac{35}{12}$$

To determine the covariance $\mathrm{W}(X, Y)$ first evaluate $\mathrm{E}(XY)$:

$$\mathrm{E}(XY) = \sum_{r=0}^{6} \sum_{s=0}^{6} r.s\,\mathrm{P}(X = r, Y = S)$$

There are only six non-zero probabilities, all $\frac{1}{6}$, and hence only six terms to sum:

$$\mathrm{E}(XY) = (1.1 + 2.2 + 3.3 + 4.4 + 5.5 + 6.6)\,\frac{1}{6} = \frac{1 + 4 + 9 + 16 + 25 + 36}{6} = \frac{91}{6}$$

The covariance can now be determined:

$$\mathrm{W}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\,.\,\mathrm{E}(Y) = \frac{91}{6} - \frac{7}{2} \times \frac{7}{2} = \frac{91}{6} - \frac{49}{4} = \frac{35}{12}$$

Accordingly, the correlation coefficient is:

$$\mathrm{R} = \frac{\mathrm{W}(X, Y)}{\sqrt{\mathrm{V}(X)\,.\,\mathrm{V}(Y)}} = \frac{\frac{35}{12}}{\sqrt{\frac{35}{12} \cdot \frac{35}{12}}} = 1$$

This is complete positive correlation.

Example — Two Reverse-Linked Dice

Suppose the two dice behave so as to give only the (equiprobable) pairs 16, 25, 34, 43, 52 and 61. The analysis is as before except that sum now is:

$$\mathrm{E}(XY) = (1.6 + 2.5 + 3.4 + 4.3 + 5.2 + 6.1)\,\frac{1}{6} = \frac{6 + 10 + 12 + 12 + 10 + 6}{6} = \frac{56}{6}$$

$$\mathrm{W}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\,.\,\mathrm{E}(Y) = \frac{56}{6} - \frac{49}{4} = -\frac{35}{12}$$

Accordingly, the correlation coefficient is:

$$\mathrm{R} = \frac{\mathrm{W}(X, Y)}{\sqrt{\mathrm{V}(X)\,.\,\mathrm{V}(Y)}} = \frac{-\frac{35}{12}}{\sqrt{\frac{35}{12} \cdot \frac{35}{12}}} = -1$$

This is complete negative correlation.

Example — A Curious Special Case

Suppose the two random variables X and Y give rise to the set of elementary events whose probabilities are as shown in the following 2×3 table:

			Y			
		s	$\rightarrow$			
			0	1	2	
X	r	0	$\frac{1}{3}$	0	$\frac{1}{3}$	$\frac{2}{3}$
	$\downarrow$	1	0	$\frac{1}{3}$	0	$\frac{1}{3}$
			$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	

The expectations and variances of X and Y are easily computed as:

$$\mathrm{E}(X) = \frac{1}{3} \qquad \mathrm{E}(Y) = 1 \qquad \mathrm{V}(X) = \frac{2}{9} \qquad \mathrm{V}(Y) = \frac{2}{3}$$

Evaluate $\mathrm{E}(XY)$:

$$\mathrm{E}(XY) = \sum_{r=0}^{1} \sum_{s=0}^{2} r.s \, \mathrm{P}(X = r, Y = S)$$

There are six terms:

$$\mathrm{E}(XY) = (0.0).\frac{1}{3} + (0.1).0 + (0.2).\frac{1}{3} + (1.0).0 + (1.1).\frac{1}{3} + (1.2).0 = \frac{1}{3}$$

Evaluate the covariance:

$$\mathrm{W}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X) \, . \, \mathrm{E}(Y) = \frac{1}{3} - \frac{1}{3} \times 1 = 0$$

The covariance is zero so the correlation is zero too. This result though is *not* obvious from the start because the random variables are clearly not independent. By inspection, each elementary event has probability 0 or $\frac{1}{3}$ but the six products $\mathrm{P}(X = r) \, . \, \mathrm{P}(X = r)$ are all either $\frac{2}{9}$ or $\frac{1}{9}$.

Footnote about Correlation

If the variables are independent the correlation is zero but...

If the correlation is zero the variables are *not* necessarily independent.

One Random Variable *versus* Two

Given a single random variable X which has some value r ($r \in \mathbb{N}$), three entities are of immediate interest:

I	$\mathrm{P}(X = r)$	The associated set of probabilities
II	$\mathrm{E}(X)$	The expectation
III	$\mathrm{V}(X)$	The variance

If X now appears in conjunction with a second random variable Y which has some value s ($s \in \mathbb{N}$), three further entities may be of interest:

I	$\mathrm{P}(X = r, Y = s)$	The set of probabilities of the composite elementary events
II	$\mathrm{E}(X + Y)$	The expectation of the sum
III	$\mathrm{V}(X + Y)$	The variance of the sum

The only reason for singling out the expectation and variance of the *sum* (rather than, say, the product) of X and Y is that the formulae for deriving these values have been discussed at length.

Concentrating on the sum for a moment, there is one entry which is conspicuous by its absence from the list. This is the *probability* of the sum. Before such an entry can be added, some meaning has to be given to the concept. If you throw two dice and obtain two values r and s you can trivially note their sum. This is a derived random variable.

It is then perfectly reasonable to ask what the probability is of this sum being, say, 7. One might express the probability as $\mathrm{P}(X + Y = t)$ where t is the sum from some particular outcome.

Here are some preliminary observations:

- Determining the *expectation* of the sum, $\mathrm{E}(X + Y)$, is easy. This is simply the sum of the expectations. It is not necessary for X and Y to be independent and they don't have to be identically distributed.
- Determining the *variance* of the sum, $\mathrm{V}(X + Y)$, is not quite so easy unless X and Y satisfy the restriction that they are independent. The variance of the sum is then sum of the variances. The variables do not have to be identically distributed.
- Determining the *probability* of the sum, $\mathrm{P}(X + Y = t)$, is difficult and will be discussed further. Certainly no progress can be made by simply adding probabilities whatever restrictions are imposed.

The Two-Dice Example

The following table shows the probabilities associated with a fair die:

	$r \rightarrow$						
X	0	1	2	3	4	5	6
$\mathrm{P}(X = r)$	0	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Here X is the random variable and $\mathrm{P}(X = r)$ is the probability of throwing value r. Take a second fair die with an associated random variable Y; the probabilities $\mathrm{P}(Y = s)$ are as for X.

Assuming independence, the 36 pairs which do not involve zero are all equiprobable with the probability for each being $\frac{1}{36}$.

A simple approach to investigating the probabilities of the different sums of the two dice scores is to consider each possible sum in turn. It is impossible to score a sum of 0 or 1 since at least one die would have to show zero and $\mathrm{P}(X = 0) = \mathrm{P}(Y = 0) = 0$. The minimum sum is 2, which is obtained by throwing two 1s.

So far, $\mathrm{P}(X + Y = 0) = 0$, $\mathrm{P}(X + Y = 1) = 0$ and $\mathrm{P}(X + Y = 2) = \frac{1}{36}$ (note that there is only one way of throwing two 1s).

A sum of 3 can be obtained in two ways, $1 + 2$ and $2 + 1$ and so $\mathrm{P}(X + Y = 2) = \frac{2}{36}$. It is straightforward to consider the other possible sums and obtain the following table:

							$t \rightarrow$						
$X + Y$	0	1	2	3	4	5	6	7	8	9	10	11	12
$\mathrm{P}(X + Y = t)$	0	0	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Note that the most-probable sum is 7 since this may be obtained in six ways: $1 + 6$, $2 + 5$, $3 + 4$, $4 + 3$, $5 + 2$ and $6 + 1$.

This duly achieves the goal of determining the values of $\mathrm{P}(X + Y = t)$ but it has been fairly hard work even in this simple case! This table provides an opportunity for a direct illustration of the formulae $\mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y)$ and $\mathrm{V}(X + Y) = \mathrm{V}(X) + \mathrm{V}(Y)$ (when X and Y are independent)...

$$\begin{aligned}\mathrm{E}(X+Y) &= \sum_{t=0}^{12} t\,\mathrm{P}(X+Y=t)\\ &= (0.0+1.0+2.1+3.2+4.3+5.4+6.5+7.6+8.5+9.4+10.3+11.2+12.1)\,\frac{1}{36}\\ &= \frac{252}{36} = 7 = \mathrm{E}(X)+\mathrm{E}(Y) \quad \left[\frac{7}{2}+\frac{7}{2}=7\right]\end{aligned}$$

$$\begin{aligned}\mathrm{E}\big((X+Y)^2\big) &= \sum_{t=0}^{12} t^2\,\mathrm{P}(X+Y=t)\\ &= (0.0+1.0+4.1+9.2+16.3+25.4+36.5+49.6+\\ &\qquad 64.5+81.4+100.3+121.2+144.1)\,\frac{1}{36} = \frac{1974}{36} = \frac{329}{6}\end{aligned}$$

These two results are used for the variance:

$$\mathrm{V}(X+Y) = \mathrm{E}\big((X+Y)^2\big) - \big(\mathrm{E}(X+Y)\big)^2 = \frac{329}{6} - 49 = \frac{35}{6} = \mathrm{V}(X)+\mathrm{V}(Y) \quad \left[\frac{35}{12}+\frac{35}{12}=\frac{35}{6}\right]$$

A Polynomial with Probabilities as Coefficients

Determining the table of values of $\mathrm{P}(X+Y=t)$ is tedious and fortunately there is an alternative approach...

Consider the following function $G(\eta)$ which is a sixth-order polynomial whose coefficients are the probabilities associated with a fair die:

$$G(\eta) = 0\eta^0 + \frac{1}{6}\eta^1 + \frac{1}{6}\eta^2 + \frac{1}{6}\eta^3 + \frac{1}{6}\eta^4 + \frac{1}{6}\eta^5 + \frac{1}{6}\eta^6$$

The coefficients could be taken straight from the table giving values of $\mathrm{P}(X=r)$ or $\mathrm{P}(Y=s)$.

What is interesting is to note the outcome of multiplying this polynomial by itself:

$$\begin{aligned} G(\eta)\,.\,G(\eta) = 0\eta^0 + 0\eta^1 + \frac{1}{36}\eta^2 + \frac{2}{36}\eta^3 + \frac{3}{36}\eta^4 + \frac{4}{36}\eta^5 + \frac{5}{36}\eta^6 + \frac{6}{36}\eta^7 \\ + \frac{5}{36}\eta^8 + \frac{4}{36}\eta^9 + \frac{3}{36}\eta^{10} + \frac{2}{36}\eta^{11} + \frac{1}{36}\eta^{12} \end{aligned}$$

The coefficients in this product polynomial are exactly those in the table of values for $\mathrm{P}(X+Y=t)$.

There will be more to say about functions like $G(\eta)$.

Glossary

The following technical terms have been introduced:

Independent and Identically Distributed, IID Correlation Coefficient

Exercises — V

Work in fractions.

1. The *Problem of Points* is the 'founding problem' of probability theory. It was discussed by Pascal and Fermat in a famous correspondence in the summer of 1654. A and B stake equal money on winning a simple coin-tossing game in which the winner of each toss scores a point, the first to reach an agreed number of points winning the game. However, the game is interrupted when A still needs two points to win, and B three (Pascal's actual example). In what proportion should the total stake be divided and returned to the players? Display the sample space on a lattice diagram, or otherwise, and find the probabilities, and hence the answer, by enumeration.

2. Determine the values of the probabilities $\mathrm{P}(X+Y=t)$ when (a) two dice are linked so $r=s$ and (b) two dice are reverse-linked so $r+s=7$. Follow the analysis on page 5.7 and directly evaluate the expectations of $X+Y$ and $(X+Y)^2$ then, from these expectations, determine the variance of $X+Y$. Note that $\mathrm{E}(X+Y)$ is, in both cases, the same as for independent dice and that $\mathrm{V}(X+Y)$, in both cases, is not.

3. Suppose X is a random variable whose value is the outcome of throwing a fair die and suppose Y is a random variable whose value is the number of heads which show when 4 fair coins are tossed. Tabulate the values of the probabilities $\mathrm{P}(X+Y=t)$ and directly evaluate the expectations of $X+Y$ and $(X+Y)^2$ then, from these expectations, determine the variance of $X+Y$. Check that $\mathrm{E}(X+Y)=\mathrm{E}(X)+\mathrm{E}(Y)$ and $\mathrm{V}(X+Y)=\mathrm{V}(X)+\mathrm{V}(Y)$.

4. Write an ML function `dice` of type `int -> int list * int` which is used thus:

```
- dice 1;
> ([0,1,1,1,1,1,1],6) : int list * int
-
- dice 2;
> ([0,0,1,2,3,4,5,6,5,4,3,2,1],36) : int list * int
```

   The function takes an `int` argument `n` and evaluates the probabilities associated with the sum of the scores when `n` fair dice are thrown. The result is a two-tuple whose first component is an `int list` of the numerators of the probabilities and whose second component is an `int` being the denominator common to all the probabilities.

   Hint: it may be helpful to note that one way of generating the list in `dice 2` is to sum the following lists, element by element:

```
[0,0,0,0,0,0,0,1,1,1,1,1,1]
[0,0,0,0,0,0,1,1,1,1,1,1]
[0,0,0,0,0,1,1,1,1,1,1]
[0,0,0,0,1,1,1,1,1,1]
[0,0,0,1,1,1,1,1,1]
[0,0,1,1,1,1,1,1]
```