Exercises 2
Relationships between events illustrated.

Two screening test are available for identifying loan defaulters. These tests are used to evaluate requests for new loans.

Test A – Credit rating result: good (G) or poor (P)

Test B - Employment Status result : satisfactory (S) or not satisfactory (NS)

Statistical analysis has provided the following estimates of the test performances produced the following tables:

Sample size = 10158 loans

Default = D, non default = $\overline{D}$

Test A	D	$\overline{D}$
P	135	1008
G	255	8760

Test B	D	$\overline{D}$
NS	120	444
S	270	9324

We shall estimate the probabilities from this large sample.

Dividing the cells of A by the sample size we get the following probabilities:

Test A	D	$\overline{D}$
P	0.01329	0.099232
G	0.025103	0.862374

What we have in the table is that:

$P((A = P) \cap D) = 0.01329$ etc

$P(D) = 0.01329 + 0.025103 = 0.038393$ or nearly 4% of loans default.

Accuracy of the test at picking up potential defaulters is described by:

$$P(D \mid P) = \frac{P(D \cap P)}{P(P)} = \frac{0.01329}{0.01329 + 0.099232} = 0.11811$$

Thus 12% of people with poor credit rating default, in 88% of these cases the loan is OK.

Suppose the evaluating software applies this criterion rejecting those with the result P.

$$P(D \mid G) = \frac{P(D \cap G)}{P(G)} = \frac{0.025103}{0.025103 + 0.862374} = 0.028286$$

A reduction in the bad credit rate from 3.8% to 2.8%. However customers will be lost:

$P(P) = 0.01329 + 0.099232 = 0.112522$, 11.3% of customers are rejected.

$P(\overline{D} \mid P) = \frac{0.099232}{0.112522} = 0.88199$, 88.2% of these would have been OK.

Suppose an OK loan nets the company $2000, the loss on a defaulted loan is $10000.

If they don't use the criterion they will make on (large) N loan applications:

N * (0.038393* (-10000) + 0.961607*2000)=1539.279*N

If they do then

N*0.887478*(0.028286*(-10000)+0.971714*2000)=1473.715*N

So they should leave well alone or get better info they are rejecting too many.

Exercises:

As exercise evaluate Test B in the same way.

If you could only use one test which would you use?

Applying two tests – assume conditional indpendence :

Suppose

$$P(T_1 \cap T_2 \mid D) = P(T_1 \mid D) \times P(T_2 \mid D)$$

and

$$P(T_1 \cap T_2 \mid \overline{D}) = P(T_1 \mid \overline{D}) \times P(T_2 \mid \overline{D})$$

where $T_1 = G$ or $T_1 = P$, $T_2 = S$ or $T_2 = NS$.

If the following procedure is adopted:

Loan approved only if T_1=G and $T_2 = S$ - what is the probability of a default?
What is the probability of rejecting a non-defaulter?

Using both the tests.

We now have 3 events (P,G), (S,NS) and (D, $\overline{D}$)

The two tables above do not (in general) contain enough information for the 3 event model.
What is the relationship between the tests?

$P(P \cap S) = ?$
How do the tests jointly impact on P(D) ?

The full model involves 7 parameters which can be arranged in the following table:

		Defaulting	
Test A	Test B	D	$\overline{D}$
P	NS	108	432
	S	27	576
G	NS	12	12
	S	243	8748

Note: the data in this table is the disaggregated version of the data in the tables above.

So that $P(D) = 0.038393$ is the same as before.

We can get the probabilities as:

$$P(P \cap NS \cap D) = \frac{108}{10158} \quad \text{etc.}$$

again

$$P(D \mid P \cap NS) = \frac{P(P \cap NS \cap D)}{P(P \cap NS)} = \frac{108}{108 + 432} = 0.2$$

as dividing the numerator and denominator by 10158 gives the same answer.

Thus 20% of loans which fail both tests will default. 80% won’t. We might have hoped for a better result!

$P(P \cap NS) = \dfrac{108 + 432}{10158} = 0.05316$ i.e. 5.3% of application fail both tests

$P(D \mid G \cap S) = \dfrac{243}{243 + 8748} = 0.027027$, 2.7% defaulters.

The disappointing results arise from the fact that (in this example) the test carry similar information. You are not likely to have a good credit rating if your employment status is NS.

$P(G \mid NS) = \dfrac{12 + 12}{12 + 12 + 108 + 432} = 0.04255$ i.e. only 4.3%

but note

$P(S \mid P) = \dfrac{27 + 576}{27 + 576 + 108 + 432} = 0.527559$ lots of people have a poor credit rating for reasons other than employment.

Models
In a 2x2 table (two events) of probabilities we have 3 parameters – 4 cells but these have to add to 1. A common parameterisation is:
1 parameter for the frequency of event A (as opposed to not(A))
1 parameter for the frequency of event B (as opposed to not(B))
1 parameter for association.- measuring the relationship.
As these are not probabilities and are measured on a log scale we defer the discussion to later in the course.

For a 2x2x2 table (three events) we have 7 parameters.
3 – for the frequencies of the events,
3 – for the binary associations between events
1 – for the triple association between events.