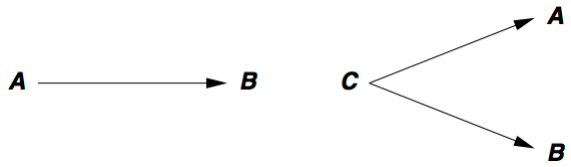
In [1]:
```
# headings↔
```

```
Populating the interactive namespace from numpy and matplotlib
```

2.4. Detection of Dependencies between variables

- Systematic Dependency between two (or more) variables is a frequent indicator for a causal relationship or a joint dependency



- A='a flash', B='a thunder' is a causal relation
- A='a wet road', B='open umbrellas' ⇒ no causal relation but dependent on a joint cause C="it rains"

This motivates our interest on 2 questions:

- how probable is the existence of a systematic dependency given the data
- if given, how strong is the dependency?

We first consider the case of nominal variables, i.e. values are in a discrete set without structure.

- Note: continuous variables can be transformed into discrete problems by binning.
- ⇒ Contingency table stores all information about the dependency of two variables.

Example: size and price of a product (e.g. laptop), sample of 170 devices

A\B	cheap	expensive	sum
small	85	17	102
big	15	53	68
sum	100	70	170

Notation:

- N_{ij} = # of cases with A = i-th value and B = j-th value.
- $N_{i*} = \sum_j N_{ij}$ row marginal, sums over columns.
- $N_{*j} = \sum_i N_{ij}$ column marginal, sums over rows.
- $N_{**} = \sum_{ij} N_{ij} = N$ total number of data points

Null hypothesis: the variables A and B are independent.

Under the assumption of a valid null hypothesis, we expect $N_{ij} = \tilde{N_{ij}}$ with

$$\frac{\tilde{N}_{ij}}{N} = \frac{N_{i*}}{N} \cdot \frac{N_{*j}}{N}$$

- Note that the condition $P(E \cap F) = P(E) \cdot P(F)$ holds for statistical independence of events E and F.
- The probabilities are estimated by the relative frequencies
- Consider the 'hit number' N_{ij} as a random variable
- The variance σ_{ij}^2 can be estimated as:

$$\sigma_{ij}^2 = \frac{\tilde{N}_{ij}}{N} \cdot \left(1 - \frac{\tilde{N}_{ij}}{N}\right) \cdot N \approx \tilde{N}_{ij}$$

- Note that the random variable can - again - be regarded as a Bernoulli process with hit probability $p = N_{ij}/N$
- the variance in this case is $\sigma^2 = npq$ with $q = 1 - p$. With the assumption that $q \approx 1$ follows the above statement.

$$\Rightarrow \left(\frac{(N_{ij} - \tilde{N}_{ij})}{\sqrt{\tilde{N}_{ij}}}\right)$$

are random variables with mean 0 and variance 1 (assuming that H_0 holds)

$$\Rightarrow \chi^2 = \sum_{ij} \frac{(N_{ij} - \tilde{N}_{ij})^2}{\tilde{N}_{ij}}$$

is approximately χ^2-distributed

- Note: because for large N_{ij} the Bernoulli distribution converges against the normal distribution
  - with $\nu = I \cdot J - I - J + 1$ degrees of freedoms, where $I = \text{number of rows and } J = \text{number of columns}$ of the contingency table.
  - Note that the number of summands is $I \cdot J$, minus the number of boundary conditions for row- and column marginals I and J. However, this subtracts one too many as the sum of the marginals is equal $\sum N_{\star j} = \sum N_{i\star} = N$.
  - With that, for a given critical value χ_c^2 the probability of error is:

$$P(\chi^2 > \chi_c^2) = Q(\chi_c^2, \nu) = \int_{\chi^2}^{\infty} p(\chi^2) d(\chi^2)$$

  - $\rightarrow$ the same decision procedure as with the χ^2-test.

Question: If the null hypothesis can be rejected, how strong is the dependency?

- Reparameterization, so that the value is independent of I and J.

The strength is essentially given by the value of χ^2, conventional normalizations are:

2.4.1 Cramer's V:

$$V := \sqrt{\frac{\chi^2}{N \cdot \min(I-1, J-1)}} \quad \in [0,1]$$

where

- $V = 1$ $\Leftarrow$ perfect Association (i.e. one variable determines the other)
- $V = 0$ $\Leftarrow$ no association at all

If $I = J = 2$, V is called 'ϕ-Statistic'.

- symmetric: does not depend on row/column choice, nor permutation of rows/columns.
- Denominator = info on table dimension $\rightarrow$ corrects for the problem that measures of association for tables of different dimension are difficult to compare directly.
- used if $I \neq J$

2.4.2 Contingency coefficient:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} \in [0,1[$$

- The value $C = 1$ will never be reached
- Is only useful to compare the strengths of association of tables with equal (I, J)
- Both with V and C there is no direct statistical interpretation of values in between
  - For instance: if you observe an association between iris color of the groupier and the roulette color amounting to $V = 0.028$, you can not derive from that value whether the association is strong enough to earn money on behalf of that.

- Recommended reading:
  - Cramers V: http://faculty.vassar.edu/lowry/newcs.html (http://faculty.vassar.edu/lowry/newcs.html)
  - Contingency coefficient: https://en.wikipedia.org/wiki/Contingency_table (https://en.wikipedia.org/wiki/Contingency_table)

[ws2017EOT20171108]