Data Mining 2018
Graphical Models for Discrete Data
Part 1: Undirected Graphs

Ad Feelders

Universiteit Utrecht

September 26, 2018

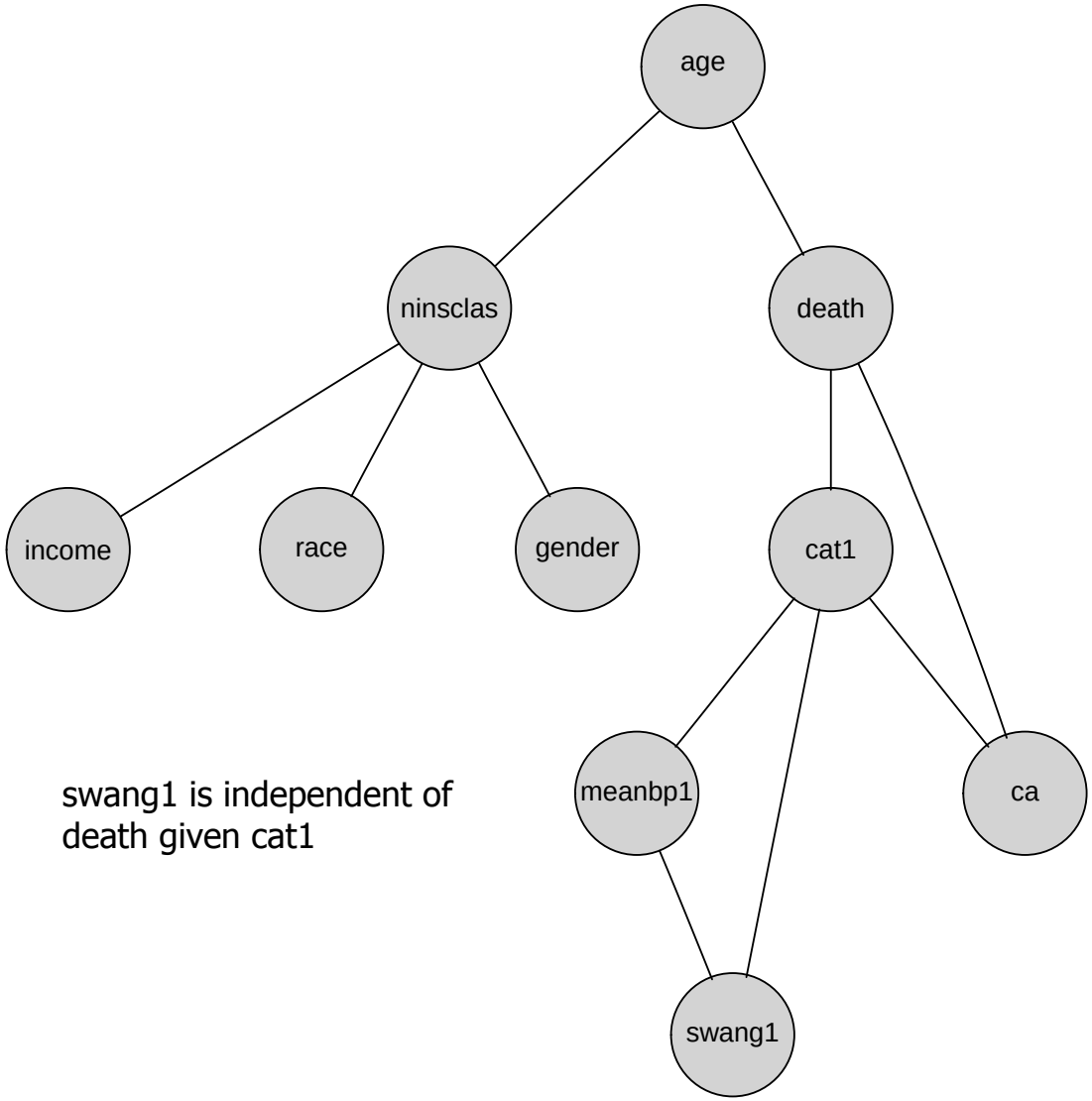
Overview of Coming Two Lectures

- (Conditional) Independence
- Graphical Representation
- Log-linear Models
  - Hierarchical
  - Graphical
  - Decomposable
- Maximum Likelihood Estimation
- Model Testing
- Model Selection (Data Mining)

Graphical Models for Discrete Data

- Task: model the associations (dependencies) between a collection of discrete variables.
- There is no designated *target* variable to be predicted: all variables are treated equal.
- This doesn't the model can't be used for prediction. It can!

Graphical Model for Right Heart Catheterization Data



swang1 is independent of death given cat1

An example

Consider the following table of counts on X and Y:

$n(x,y)$	y			
x	q	r	s	$n(x)$
a	2	5	3	10
b	10	20	10	40
c	8	35	7	50
$n(y)$	20	60	20	100

Suppose we want to estimate the joint distribution of X and Y.

The Saturated Model

Saturated Model

$$\hat{P}(x,y) = \frac{n(x,y)}{n}$$

requires the estimation of 8 probabilities.

The fitted counts $\hat{n}(x,y) = n\hat{P}(x,y)$ are the same as the observed counts.

$\hat{n}(x,y)$	y			
x	q	r	s	$n(x)$
a	2	5	3	10
b	10	20	10	40
c	8	35	7	50
$n(y)$	20	60	20	100

The Saturated Model and the Curse of Dimensionality

The saturated model estimates cell probabilities by dividing the cell count by the total number of observations. It makes no simplifying assumptions. This approach doesn't scale very well!

Suppose we have k categorical variables with m possible values each.

To estimate the probability of each possible combination of values would require the estimation of m^k probabilities. For $k = 10$ and $m = 5$, this is

$$5^{10} \approx 10 \text{ million probabilities}$$

This is a manifestation of the *curse of dimensionality*: we have fewer data points than probabilities to estimate.

How to avoid this curse

Look for appropriate independence assumptions to find a simpler model that still gives a good fit.

Independence Model

$$\hat{P}(x,y) = \hat{P}(x)\hat{P}(y) = \frac{n(x)}{n}\frac{n(y)}{n} = \frac{n(x)n(y)}{n^2}$$

requires the estimation of just 4 probabilities.

Fit of independence model

The fitted counts of the independence model are given by

$$\hat{n}(x,y) = n\hat{P}(x,y) = n\,\frac{n(x)n(y)}{n^2} = \frac{n(x)n(y)}{n}$$

Compare the fitted counts with the observed counts:

$\hat{n}(x,y)$	y			
x	q	r	s	$\hat{n}(x)$
a	2	6	2	10
b	8	24	8	40
c	10	30	10	50
$\hat{n}(y)$	20	60	20	100

$n(x,y)$	y			
x	q	r	s	$n(x)$
a	2	5	3	10
b	10	20	10	40
c	8	35	7	50
$n(y)$	20	60	20	100

Fit of independence model

- The fitted counts of the independence model are quite close to the observed counts.
- We could conclude that the independence model gives a satisfactory fit of the data.
- Use a statistical test to make this more precise (discussed later).

Independence Model

- The saturated model requires the estimation of $m^k - 1$ probabilities.
- The mutual independence model requires just $k(m-1)$ probability estimates.
- Mutual independence model is usually not appropriate (all variables are independent of one another).
- Interesting models are somewhere in between saturated and mutual independence: this requires the notion of *conditional* independence.

Rules of Probability

1. Sum Rule:
$$P(X) = \sum_{Y} P(X, Y)$$
2. Product Rule:
$$P(X, Y) = P(X)P(Y|X)$$
3. If X and Y are independent, then
$$P(X, Y) = P(X)P(Y)$$

Independence of (sets of) random variables

Let X and Y be (sets of) random variables.
X and Y are independent if and only if:

$$P(x, y) = P(x)P(y) \text{ for all values } (x, y).$$

As a consequence

$$P(x \mid y) = P(x), \text{ and } P(y \mid x) = P(y)$$

Y doesn't provide any information about X (and vice versa)

We also write $X \perp\!\!\!\perp Y$.

For example: gender is independent of eye color.

Factorisation criterion for independence

We can relax our burden of proof a little bit:

X and Y are independent iff there are functions $g(x)$ and $h(y)$ (not necessarily the marginal distributions of X and Y) such that

$$P(x, y) = g(x)h(y)$$

In logarithmic form this becomes (since $\log ab = \log a + \log b$):

$$\log P(x, y) = g^*(x) + h^*(y),$$

where $g^*(x) = \log g(x)$.

Factorisation criterion for independence: proof

Suppose that for all x and y:

$$P(x, y) = g(x)h(y)$$

Then

$$P(x) = \sum_y P(x, y) = \sum_y g(x)h(y) = g(x) \sum_y h(y) = c_1 \, g(x)$$

So $g(x)$ is proportional to $P(x)$. Likewise, $h(y)$ is proportional to $P(y)$. Therefore

$$P(x, y) = g(x)h(y) = \frac{1}{c_1} P(x) \frac{1}{c_2} P(y) = c_3 P(x)P(y)$$

Summing over both x and y establishes that $c_3 = 1$, so X and Y are independent.

Conditional Independence

X and Y are *conditionally* independent given Z iff

$$P(x, y \mid z) = P(x \mid z)P(y \mid z)$$

for all values (x, y) and for all values z for which $P(z) > 0$. Equivalently:

$$P(x \mid y, z) = P(x \mid z)$$

If I know the value of Z, then Y doesn't provide any additional information about X.

We also write $X \perp\!\!\!\perp Y \mid Z$.

For example: ice cream sales is independent of mortality among the elderly given the weather.

Factorisation Criterion for Conditional Independence

An equivalent formulation is

$$P(x, y, z) = \frac{P(x, z)P(y, z)}{P(z)}$$

Factorisation criterion: $X \perp\!\!\!\perp Y \mid Z$ iff there exist functions g and h such that

$$P(x, y, z) = g(x, z)h(y, z)$$

or alternatively

$$\log P(x, y, z) = g^*(x, z) + h^*(y, z)$$

for all (x, y) and for all z for which $P(z) > 0$.

Conditional Independence Graph

Random Vector $X = (X_1, X_2, \ldots, X_k)$ with probability distribution $P(X)$.
Graph $G = (K, E)$, with $K = \{1, 2, \ldots, k\}$.

The conditional independence graph of X is the undirected graph $G = (K, E)$ where $\{i, j\}$ is *not* in the edge set E if and only if:

$$X_i \perp\!\!\!\perp X_j \mid \text{rest}$$

Conditional Independence Graph: Example

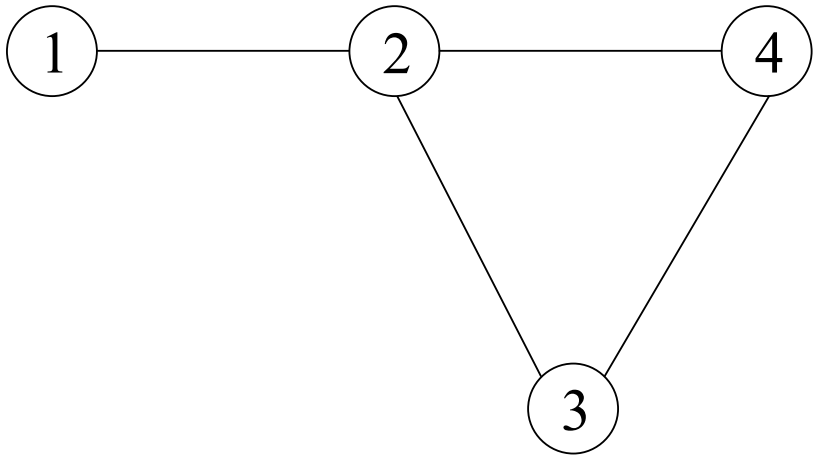
$X = (X_1, X_2, X_3, X_4), 0 < x_i < 1$ with probability density

$$P(x) = e^{c+x_1+x_1x_2+x_2x_3x_4}$$

Application of the factorisation criterion gives

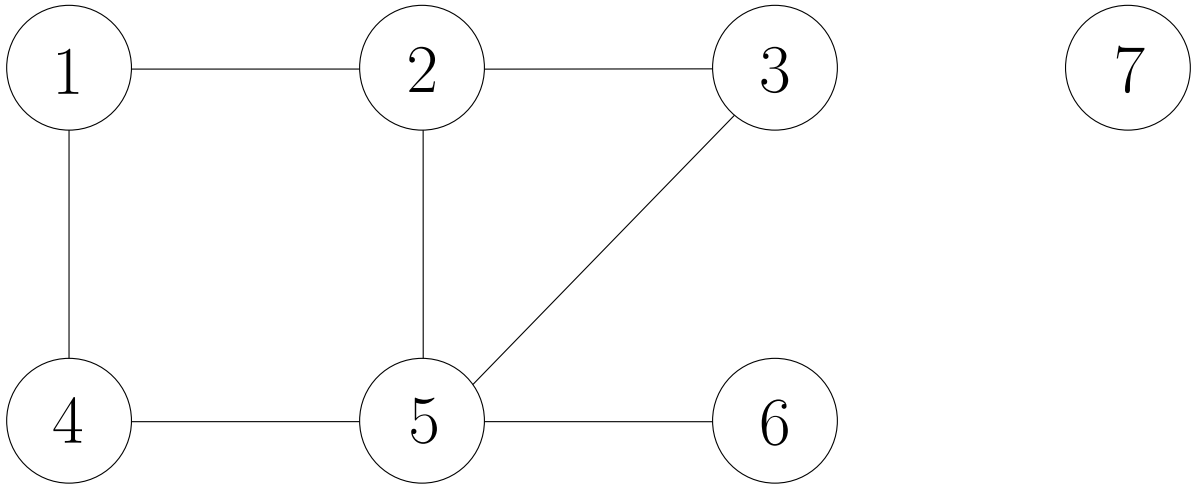
$$X_1 \perp\!\!\!\perp X_4 \mid (X_2, X_3) \text{ and } X_1 \perp\!\!\!\perp X_3 \mid (X_2, X_4),$$

so the conditional independence graph is:



Separation

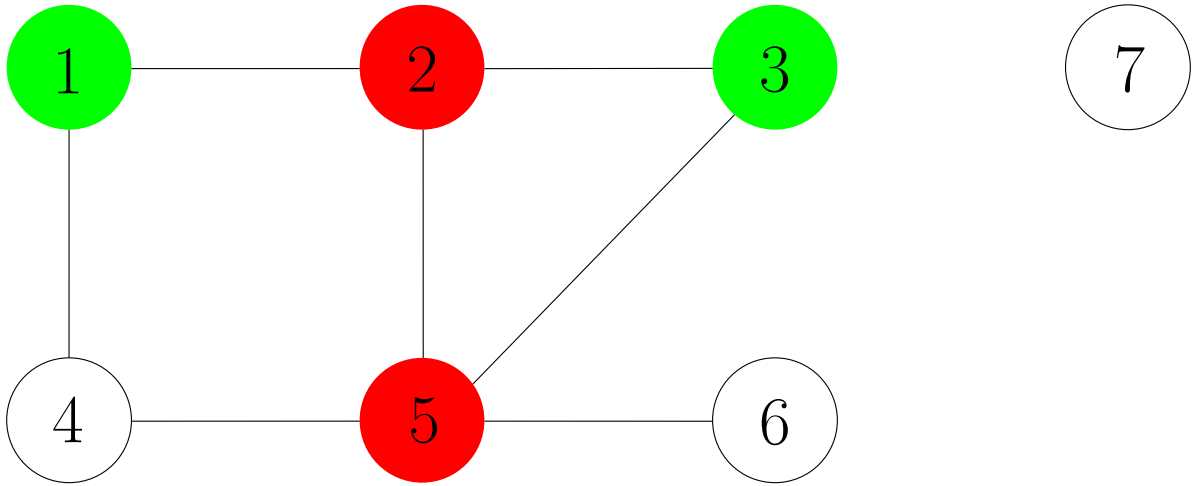
Consider the following conditional independence graph:



- $X_1 \perp\!\!\!\perp X_3 \mid (X_2, X_4, X_5, X_6, X_7)$

$\{2, 5\}$ separates 1 from 3

Consider the following conditional independence graph:



- $X_1 \perp\!\!\!\perp X_3 \mid (X_2, X_4, X_5, X_6, X_7)$
- $\{2, 5\}$ separates 1 from 3 $\Rightarrow X_1 \perp\!\!\!\perp X_3 \mid (X_2, X_5)$

Separation

Notation:

$$X_a = (X_i \ : \ i \in a)$$

where a is a subset of $\{1, 2, \ldots, k\}$.

The set a separates node i from node j iff every path from node i to node j contains one or more nodes in a.

a separates b from c (a, b, c disjoint):

$$\text{For every } i \in b \text{ and } j \in c : a \text{ separates } i \text{ from } j$$

Equivalent Markov Properties

1. Pairwise: for all non-adjacent vertices i and j

$$X_i \perp\!\!\!\perp X_j \mid \text{rest}$$

   This is how we defined the graph.
2. Global: if a separates b from c (a, b, c disjoint), then

$$X_b \perp\!\!\!\perp X_c \mid X_a$$

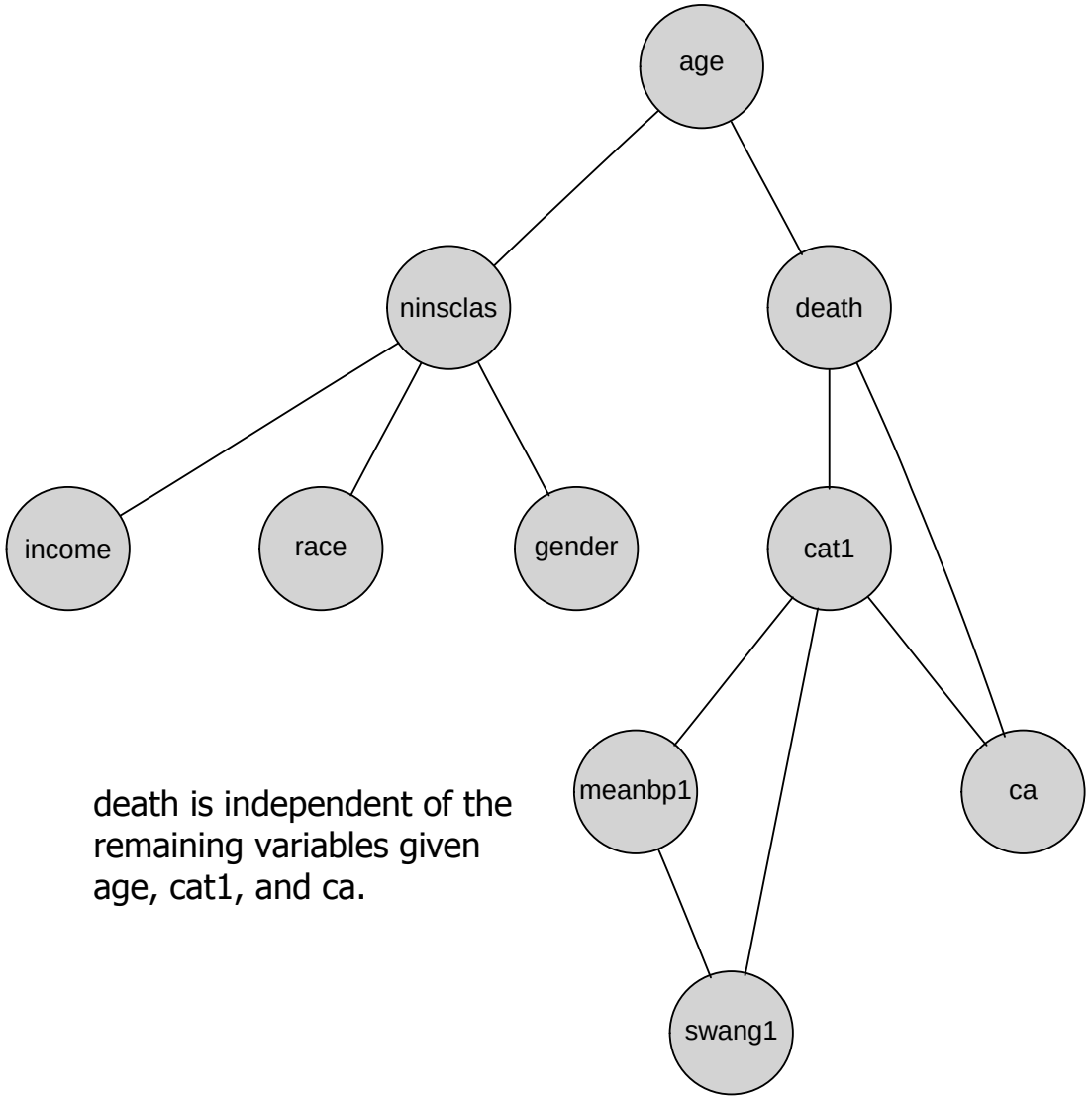
3. Local:

$$X_i \perp\!\!\!\perp \text{ rest} \mid \text{boundary}(i),$$

   where boundary(i) is the set of nodes adjacent (directly connected) to node i.

If all pairwise independencies corresponding to graph G hold for a given probability distribution, then all the global independencies corresponding to G also hold for that distribution (and vice versa).

Graphical Model for Right Heart Catheterization Data



death is independent of the remaining variables given age, cat1, and ca.

A Famous Example

Data on the survival of 715 infants attending two clinics and the amount of care received by the mother.

Table of counts for clinic, care and survival:

n(clinic, care, survival)		survival	
clinic	care	no	yes
clinic 1	less	3	176
	more	4	293
clinic 2	less	17	197
	more	2	23

A Famous Example

Assume survival and care are independent within both clinics.

This *conditional* independence assumption corresponds to the following factorization:

$$\hat{P}(\text{care, survival} \mid \text{clinic}) = \hat{P}(\text{care} \mid \text{clinic})\hat{P}(\text{survival} \mid \text{clinic})$$

Multiplying left and right by $\hat{P}(\text{clinic})$ we get

$$\begin{aligned}\hat{P}(\text{care, survival, clinic}) &= \hat{P}(\text{care, clinic})\hat{P}(\text{survival} \mid \text{clinic}) \\ &= \frac{\hat{P}(\text{care, clinic})\hat{P}(\text{survival, clinic})}{\hat{P}(\text{clinic})}\end{aligned}$$

A Famous Example

Writing $\hat{n}$ for $n\hat{P}$ we get fitted counts (multiply left and right by n):

$$\begin{aligned}\hat{n}(\text{clinic, care, survival}) &= \frac{\hat{n}(\text{clinic, care})\hat{n}(\text{clinic, survival})}{\hat{n}(\text{clinic})} \\ &= \frac{n(\text{clinic, care})n(\text{clinic, survival})}{n(\text{clinic})}\end{aligned}$$

(The last step will be explained in more detail in the next lecture.)

Now we have an expression for the fitted counts in terms of the observed counts.

Minimal Sufficient Statistics

n(clinic, care)	care	
clinic	less	more
clinic 1	179	297
clinic 2	214	25

n(clinic, survival)	survival	
clinic	no	yes
clinic 1	7	469
clinic 2	19	220

Fitted Counts and Observed Counts

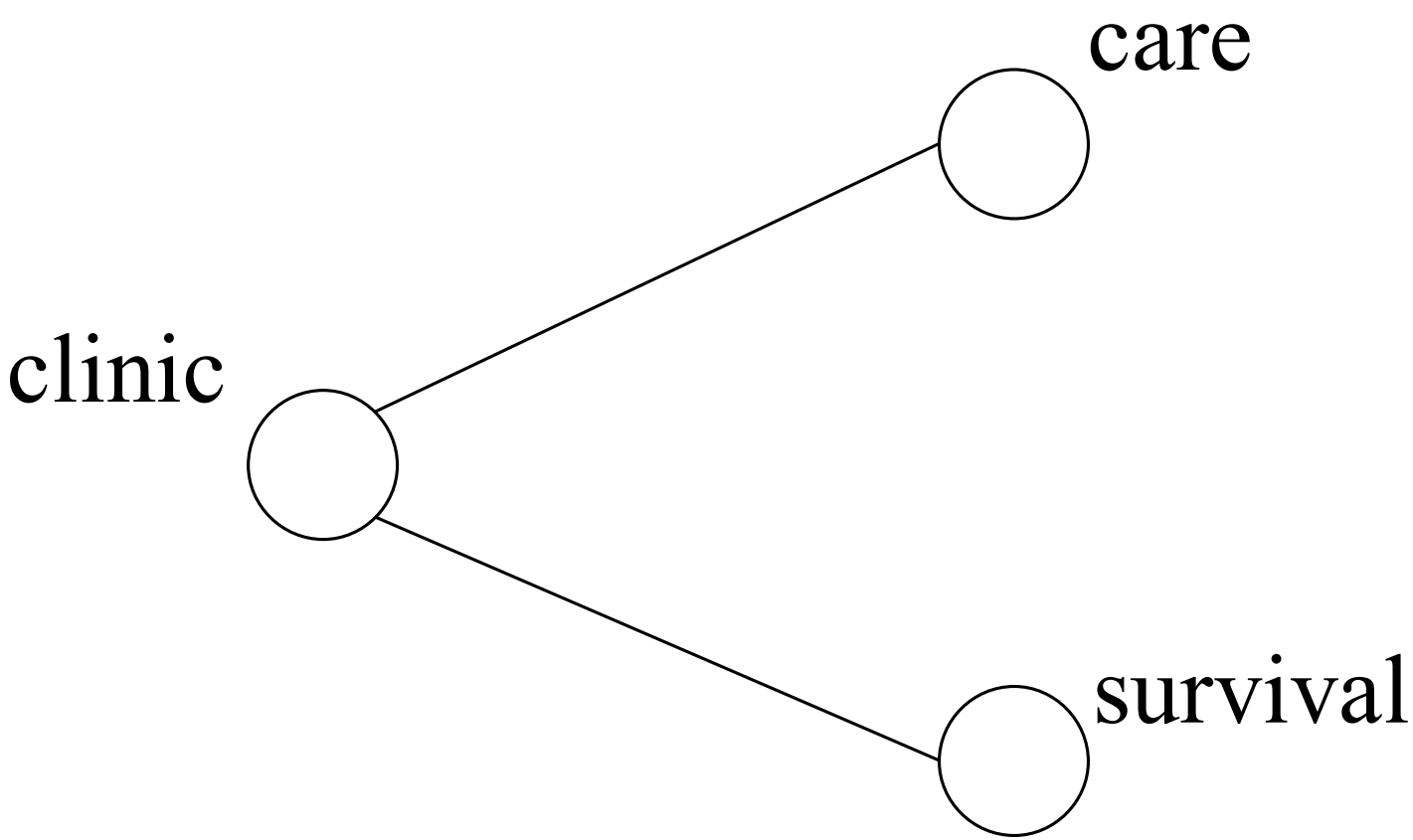
$\hat{n}$(clinic, care, survival)		survival	
clinic	care	no	yes
clinic 1	less	2.63	176.37
	more	4.37	292.63
clinic 2	less	17.01	196.99
	more	1.99	23.01

n(clinic, care, survival)		survival	
clinic	care	no	yes
clinic 1	less	3	176
	more	4	293
clinic 2	less	17	197
	more	2	23

Fitted counts are quite close to observed counts! Hence assuming care and survival are independent within both clinics seems justified.

Relation between care and survival

Graph representing conditional independence assumption:



Relation between care and survival

Summing out clinic gives:

	survival		
care	no	yes	(%)
less	20	373	5.1
more	6	316	1.9

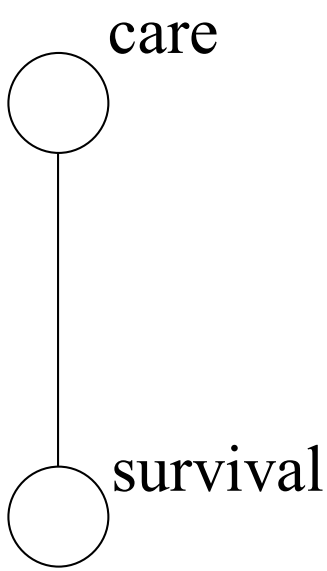
Infant mortality for mothers receiving less care is 5.1%, and for mothers receiving more care just 1.9%.
Cross-product ratio between care and survival

$$\text{cpr(care,survival)} = \frac{n(\text{less}, \text{no})n(\text{more}, \text{yes})}{n(\text{less}, \text{yes})n(\text{more}, \text{no})} = \frac{20 \times 316}{373 \times 6} = 2.82$$

But we just saw that care and survival are independent in both clinics!

Relation between care and survival

Collapsing over clinic gives the spurious association



How come?

Bernoulli random variable

Let X be a Bernoulli random variable with probability of success p, that is, $P(X = 1) = p$ and $P(X = 0) = 1 - p$.

We can write the probability function in a single formula as follows:

$$P(x) = p^x(1-p)^{1-x} \quad \text{for } x \in \{0, 1\} \text{ and } 0 \leq p \leq 1$$

Check that indeed $P(1) = p$ and $P(0) = 1 - p$ as required.

2×2 Table

The probability function P_{12} of bivariate Bernoulli random vector (X_1, X_2) is determined by

$$P(x_1, x_2) = p(x_1, x_2)$$

where $p(x_1, x_2)$ is the table of probabilities:

$p(x_1, x_2)$	$x_2 = 0$	$x_2 = 1$	Total
$x_1 = 0$	$p(0,0)$	$p(0,1)$	$p_1(0)$
$x_1 = 1$	$p(1,0)$	$p(1,1)$	$p_1(1)$
Total	$p_2(0)$	$p_2(1)$	1

Probability function for 2×2 Table

Again we can write this as one function:

$$P(x_1, x_2) = p(0,0)^{(1-x_1)(1-x_2)} p(0,1)^{(1-x_1)x_2} p(1,0)^{x_1(1-x_2)} p(1,1)^{x_1 x_2}$$

Taking logarithms and collecting terms in x_1, x_2, and $x_1 x_2$ gives:

$$\begin{aligned} \log P(x_1, x_2) &= \log p(0,0) + x_1 \log \frac{p(1,0)}{p(0,0)} + \\ & \qquad x_2 \log \frac{p(0,1)}{p(0,0)} + x_1 x_2 \log \frac{p(1,1)p(0,0)}{p(0,1)p(1,0)} \end{aligned}$$

Verify this using elementary properties of logarithms:

1. $\log a^b = b \log a$,
2. $\log \frac{a}{b} = \log a - \log b$, and
3. $\log ab = \log a + \log b$.

Log-linear expansion

Reparameterizing the right hand side leads to the so-called *log-linear expansion*

$$\log P(x_1, x_2) = u_\emptyset + u_1 x_1 + u_2 x_2 + u_{12} x_1 x_2$$

The coefficients, $u_\emptyset, u_1, u_2, u_{12}$ are known as the u-terms. For example, the coefficient of the product $x_1 x_2$,

$$u_{12} = \log \left(\frac{p(1,1)p(0,0)}{p(0,1)p(1,0)} \right) = \log \text{cpr}(X_1, X_2)$$

is the logarithm of the cross product ratio of X_1 and X_2.

Independence and u-terms

Claim:

$$X_1 \perp\!\!\!\perp X_2 \Leftrightarrow u_{12} = 0$$

Proof: the factorisation criterion states that $X_1 \perp\!\!\!\perp X_2$ iff there exist two functions g and h such that

$$\log P(x_1, x_2) = g(x_1) + h(x_2) \text{ for all } (x_1, x_2)$$

If $u_{12} = 0$, we get

$$\log P(x_1, x_2) = u_\emptyset + u_1 x_1 + u_2 x_2,$$

so

$$g(x_1) = u_\emptyset + u_1 x_1 \qquad h(x_2) = u_2 x_2$$

suffices. If $u_{12} \neq 0$, no such decomposition is possible.

Three Dimensional Bernoulli

The joint distribution of three binary variables can be written:

$$P(x_1, x_2, x_3) = p(0,0,0)^{(1-x_1)(1-x_2)(1-x_3)} \cdots p(1,1,1)^{x_1 x_2 x_3}$$

Log-linear expansion

$$\begin{aligned} \log P(x_1, x_2, x_3) &= u_\emptyset + u_1 x_1 + u_2 x_2 + u_3 x_3 + u_{12} x_1 x_2 + \\ &\quad u_{13} x_1 x_3 + u_{23} x_2 x_3 + u_{123} x_1 x_2 x_3 \end{aligned}$$

With

$$\begin{aligned} u_{123} &= \log\left(\frac{p(1,0,0)p(1,1,1)}{p(1,1,0)p(1,0,1)}\right) - \log\left(\frac{p(0,0,0)p(0,1,1)}{p(0,1,0)p(0,0,1)}\right) \\ &= \log\left(\frac{\text{cpr}(X_2, X_3|X_1 = 1)}{\text{cpr}(X_2, X_3|X_1 = 0)}\right) \end{aligned}$$

Independence and the u-terms

Observation:

$$X_2 \perp\!\!\!\perp X_3 \mid X_1 \Leftrightarrow u_{23} = 0 \text{ and } u_{123} = 0$$

Proof: use factorisation criterion.

$X_2 \perp\!\!\!\perp X_3 \mid X_1 \Leftrightarrow$ there are functions $g(x_1, x_2)$ and $h(x_1, x_3)$ such that

$$\log P(x_1, x_2, x_3) = g(x_1, x_2) + h(x_1, x_3)$$

This is only possible when $u_{23} = 0$ (so the term $x_2 x_3$ drops out), and $u_{123} = 0$ (so the term $x_1 x_2 x_3$ drops out).

Why the log-linear representation?

Why do we use the log-linear representation of the probability table?

1. We are interested in expressing conditional independence constraints.
2. There is a straightforward correspondence between such constraints being satisfied, and the elimination of certain collections of u-terms from the log-linear expansion.
3. This correspondence is established by applying the factorisation criterion: $X \perp\!\!\!\perp Y \mid Z$ if and only if there exist functions g and h such that

$$\boxed{\log P(x, y, z) = g(x, z) + h(y, z)}$$