# Data Mining
# Graphical Models for Discrete Data
# Undirected Graphs (Markov Random Fields)

Ad Feelders

Universiteit Utrecht

# Overview of Coming Lectures

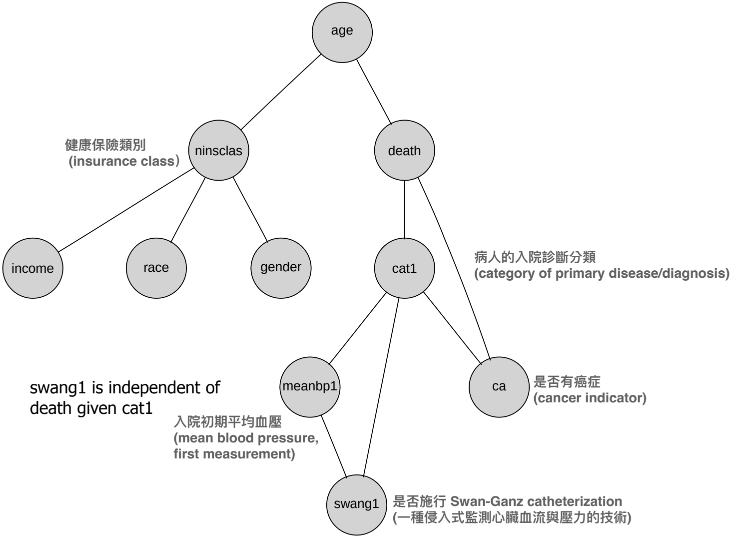
- Introduction
- Independence and Conditional Independence
- Graphical Representation of Conditional Independence
- Log-linear Models
  - Hierarchical
  - Graphical
  - Decomposable
- Maximum Likelihood Estimation
- Model Testing
- Model Selection

# Graphical Models for Discrete Data
離散的

- Task: model the associations (dependencies) between a collection of discrete variables.
- There is no designated *target* variable to be predicted: 指定 all variables are treated equal.
- This doesn't mean these models can't be used for prediction. Actually, they can, and in a more flexible way than specialized prediction models.

# Graphical Model for Right Heart Catheterization Data

導管插入




swang1 is independent of death given cat1

# Example: Gender and Eye Color

Consider the following table of counts on Gender (G) and Eye Color (E) for a random sample of $n = 100$ CS students:

| $n(G,E)$ | Eye Color | | | | |
|---|---|---|---|---|---|
| Gender | green | hazel | blue | brown | $n(G)$ |
| male | 7 | 15 | 20 | 38 | 80 |
| female | 3 | 5 | 5 | 7 | 20 |
| $n(E)$ | 10 | 20 | 25 | 45 | 100 |

Suppose we want to estimate the joint probability distribution of Gender and Eye Color.

How would you do this?

# The Saturated Model

飽和

Saturated (unconstrained) model

$$\hat{P}(G,E) = \frac{n(G,E)}{n}$$

requires the estimation of 7 probabilities.

| $\hat{P}(G,E)$ | Eye Color | | | | |
|---|---|---|---|---|---|
| Gender | green | hazel | blue | brown | $\hat{P}(G)$ |
| male | 0.07 | 0.15 | 0.20 | 0.38 | 0.8 |
| female | 0.03 | 0.05 | 0.05 | 0.07 | 0.2 |
| $\hat{P}(E)$ | 0.1 | 0.2 | 0.25 | 0.45 | 1 |

For example

$$\hat{P}(\text{male}, \text{blue}) = \frac{n(\text{male}, \text{blue})}{n} = \frac{20}{100} = 0.20.$$

# The Saturated Model

For the saturated model, the fitted counts
擬合後的計數

$$\hat{n}(G,E) = n\hat{P}(G,E) = n\left(\frac{n(G,E)}{n}\right) = n(G,E)$$

are the same as the observed counts.

| $\hat{n}(G,E)$ | Eye Color | | | | |
|---|---|---|---|---|---|
| Gender | green | hazel | blue | brown | $\hat{n}(G)$ |
| male | 7 | 15 | 20 | 38 | 80 |
| female | 3 | 5 | 5 | 7 | 20 |
| $\hat{n}(E)$ | 10 | 20 | 25 | 45 | 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.
類別變數數量 每個變數有多少個可能值

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. Estimates will become unreliable.
證實
現實中我們的樣本數 n 往往遠小於所需的組合數

# How to avoid this curse

Make independence assumptions to obtain a simpler model that still gives a good fit.

Independence Model

$$\hat{P}(G,E) = \hat{P}(G)\hat{P}(E) = \left(\frac{n(G)}{n}\right)\left(\frac{n(E)}{n}\right) = \frac{n(G)n(E)}{n^2}$$

requires the estimation of just 4 probabilities instead of 7.

1️⃣ 不假設獨立性（完整聯合機率）
聯合機率表格大小： 2×4=8 個格子
因為所有機率加起來必須等於 1，所以只需要估計 7 個獨立機率（第 8 個可以用 1 減去前面 7 個算出來）

2️⃣ 假設 G 和 E 獨立
聯合機率可以拆成：P(G,E)=P(G) · P(E)

邊際機率：
G 有 2 個值 → 只需要估計 1 個（因為 P(G 1 )+P(G 2 )=1）
E 有 4 個值 → 需要估計 3 個（因為 P(E1)+P(E2)+P(E3)+P(E4)=1
總共只要估計 1 + 3 = 4 個機率

# Fit of independence model

The fitted counts of the independence model are given by

$$\hat{n}(G,E) = n\hat{P}(G,E) = n\left(\frac{n(G)n(E)}{n^2}\right) = \frac{n(G)n(E)}{n}$$

For example

$$\hat{n}(\text{female},\text{brown}) = \frac{n(\text{female})n(\text{brown})}{n} = \frac{20 \times 45}{100} = 9$$

Table of fitted counts of the independence model:

| $\hat{n}(G,E)$ | Eye Color | | | | |
|---|---|---|---|---|---|
| Gender | green | hazel | blue | brown | $\hat{n}(G)$ |
| male | 8 | 16 | 20 | 36 | 80 |
| female | 2 | 4 | 5 | 9 | 20 |
| $\hat{n}(E)$ | 10 | 20 | 25 | 45 | 100 |

# Fit of independence model

Compare the fitted counts of the independence model (top) with the observed counts (bottom):

| $\hat{n}(G,E)$ | Eye Color | | | | |
|---|---|---|---|---|---|
| Gender | green | hazel | blue | brown | $\hat{n}(G)$ |
| male | 8 | 16 | 20 | 36 | 80 |
| female | 2 | 4 | 5 | 9 | 20 |
| $\hat{n}(E)$ | 10 | 20 | 25 | 45 | 100 |

| $n(G,E)$ | Eye Color | | | | |
|---|---|---|---|---|---|
| Gender | green | hazel | blue | brown | $n(G)$ |
| male | 7 | 15 | 20 | 38 | 80 |
| female | 3 | 5 | 5 | 7 | 20 |
| $n(E)$ | 10 | 20 | 25 | 45 | 100 |

# Fit of independence model

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

# Independence Model

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

- The saturated model requires the estimation of $m^k - 1$ probabilities.
  飽和的
- The mutual independence model requires just $k(m-1)$ probability estimates.
  不是(k-1)(m-1)
  註：(k−1)(m−1) 常出現在 二元列聯表的自由度計算，但那是針對行列數和 chi-square 檢驗的自由度，不是機率估計數量。
- Mutual independence model is usually not appropriate (all variables are independent of one another).
  假設 所有變數完全不相關
- Interesting models are somewhere in between saturated and mutual independence: this requires the notion of *conditional* independence.

# Rules of Probability

1. Sum Rule:

$$\boxed{P(X) = \sum_{Y} P(X, Y)}$$

2. Product Rule:

$$\boxed{P(X, Y) = P(X)P(Y|X)}$$

3. If $X$ and $Y$ are independent, then

$$\boxed{P(X, Y) = P(X)P(Y)}$$

# Independence of (sets of) random variables

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

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

Equivalently:

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

*$Y$ 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
因式分解

We can relax our burden of proof a little bit:
重擔

$X$ 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 $\log(a \times b) = \log a + \log b$):
對數的

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

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
呈常比的

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

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

$$P(x, y \mid z) = P(x \mid z)P(y \mid z) \tag{1}$$

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

因為 P(x,y|z)=P(x,y,z) / P(z)

Equivalently:

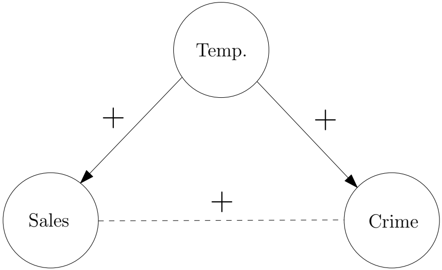
$$P(x \mid y, z) = P(x \mid z) \qquad \text{and} \qquad P(y \mid x, z) = P(y \mid z)$$

*If I already know the value of $Z$, then $Y$ doesn't provide any additional information about $X$ (and vice versa).*

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

# Conditional Independence: Example

Ice cream sales is independent of violent crime given the temperature.
極端的
Causal picture:




$\mathrm{P}(\mathrm{Crime} = \mathrm{hi} \mid \mathrm{Sales} = \mathrm{hi}) \neq \mathrm{P}(\mathrm{Crime} = \mathrm{hi})$ 如果單純比較「銷售量高 vs 犯罪率高」，兩者好像有相關。
→ 這是因為 溫度 (Temp.) 扮演了一個「潛在影響因子」

$$\mathrm{P}(\mathrm{Crime} = \mathrm{hi} \mid \mathrm{Temp.} = \mathrm{hi}, \mathrm{Sales} = \mathrm{hi}) = \mathrm{P}(\mathrm{Crime} = \mathrm{hi} \mid \mathrm{Temp.} = \mathrm{hi})$$

# Factorisation Criterion for Conditional Independence

An equivalent formulation is (multiply equation (1) by $P(z)$):

$X \perp\!\!\!\perp Y | Z \iff P(x,y|z)=P(x|z)P(y|z), \forall x,y,z$且 $P(z)>0$
兩邊同時乘上 $P(z)$

$$P(x,y,z) = \underbrace{P(x,z)}_{g(x,z)} \underbrace{\frac{P(y,z)}{P(z)}}_{h(y,z)}$$

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

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

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

# Overview of Coming Lectures

- Introduction
- Independence and Conditional Independence
- ☞ Graphical Representation of Conditional Independence
- Log-linear Models
  - Hierarchical
  - Graphical
  - Decomposable
- Maximum Likelihood Estimation
- Model Testing
- Model Selection

# 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 <u>not</u> 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

$$P(x) = e^{c + x_1 + x_1 x_2 + x_2 x_3 x_4}$$

Now

$$\log P(x) = c + x_1 + x_1 x_2 + x_2 x_3 x_4$$

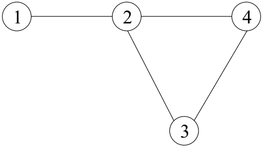
Application of the factorisation criterion gives

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

For example, $X_1 \perp\!\!\!\perp X_4 \mid (X_2, X_3)$, because we can write:

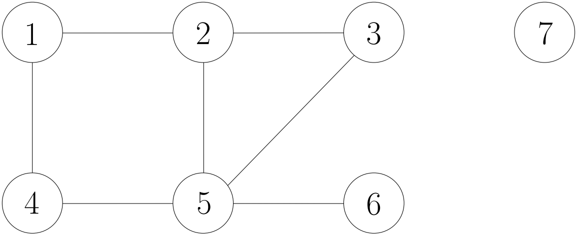
$$\log P(x) = \underbrace{c + x_1 + x_1 x_2}_{g(x_1, x_2, x_3)} + \underbrace{x_2 x_3 x_4}_{h(x_2, x_3, x_4)}$$

Hence, the conditional independence graph is:



# Separation and Conditional Independence

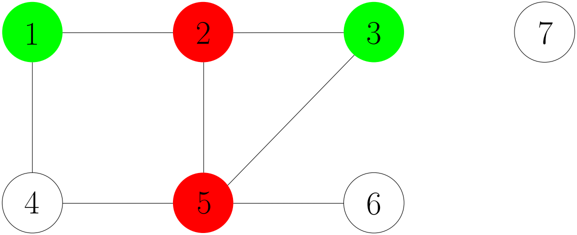
Consider the following conditional independence graph:




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

# $\{2, 5\}$ 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 $\Rightarrow X_1 \perp\!\!\!\perp X_3 \mid (X_2, X_5)$
- We also say the set $\{2, 5\}$ blocks every path from 1 to 3 (and vice versa).

# Separation and Conditional Independence

Notation:

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

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

For example, if $a = \{1, 3, 6\}$ then $X_a = (X_1, X_3, X_6)$.

The set $a$ separates node $i$ from node $j$ iff every path from node $i$ to node $j$ contains one or more nodes in $a$ ($a$ "blocks" every path from $i$ to $j$).

$a$ 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 Independence (Markov) Properties

1. Pairwise: for all non-adjacent vertices $i$ and $j$
   兩兩成對 相連 頂點

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

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

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

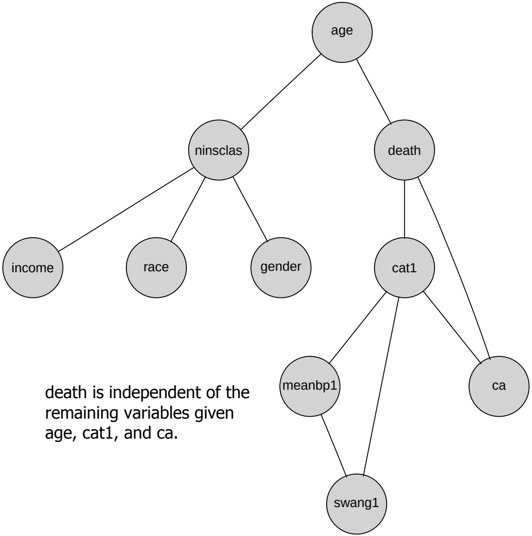
3. Local:

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

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

These properties are equivalent in the following sense: 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). Pairwise⟺Local⟺Global

# Example of Local Markov Property




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

# Overview of Coming Lectures

- Introduction
- Independence and Conditional Independence
- Graphical Representation of Conditional Independence
- ☞ Log-linear Models
    - Hierarchical
    - Graphical
    - Decomposable
- Maximum Likelihood Estimation
- Model Testing
- Model Selection

# Bernoulli random variable

Let $X$ be a Bernoulli random variable with $P(X=1)=p(1)$ and $P(X=0)=p(0)$.

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

$$P(X=x)=p(1)^x p(0)^{1-x} \qquad \text{for } x\in\{0,1\}$$

Check that filling in $x=1$ gives $p(1)$, and filling in $x=0$ gives $p(0)$ as required.

Taking logarithms we get:

$$\begin{aligned}
\log P(X=x) &= \log\left(p(1)^x p(0)^{1-x}\right)\\
&= \log p(1)^x + \log p(0)^{1-x}\\
&= x\log p(1) + (1-x)\log p(0) \qquad = \text{logp(0)+x(logp(1)–logp(0))}\\
&= \underbrace{\log p(0)}_{\text{constant}} + \underbrace{\log\frac{p(1)}{p(0)}}_{\text{coefficient of } x} \; x
\end{aligned}$$

# $2 \times 2$ 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 $2 \times 2$ Table

Again we can write this as a single formula:

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

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

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

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 $x_1 x_2$,

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

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

# Cross-product Ratio

The cross-product ratio between binary variables $X_1$ and $X_2$ is:

$$\text{cpr}(X_1, X_2) = \frac{p(1,1)p(0,0)}{p(0,1)p(1,0)}$$

- $\text{cpr}(X_1, X_2) > 1$: positive association between $X_1$ and $X_2$.
- $\text{cpr}(X_1, X_2) < 1$: negative association between $X_1$ and $X_2$.
- $\text{cpr}(X_1, X_2) = 1$: no association between $X_1$ 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

$$u_{123} = \log \left( \frac{\text{cpr}(X_2, X_3 | X_1 = 1)}{\text{cpr}(X_2, X_3 | X_1 = 0)} \right)$$

# 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

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