Markov Bases:
Hypothesis Tests for Contingency Tables

Session 4

Danai

Outline

We will talk about

- contingency tables
- statistical models
- hypothesis tests
- Markov bases.

A **contingency table** contains counts obtained by cross-classifying observed cases according to two or more discrete criteria.

	Death Penalty		
Race	Yes	No	Total
White	19	141	160
Black	17	149	166
Total	36	290	326

Classification of 326 homicide indictments in Florida in 1970s.

A **contingency table** contains counts obtained by cross-classifying observed cases according to two or more discrete criteria.

	Death Penalty		
Race	Yes	No	Total
White	19	141	160
Black	17	149	166
Total	36	290	326

Classification of 326 homicide indictments in Florida in 1970s.

Were the death penalty decisions made independently of the defendant's race?

Independent variables

Consider two random variables $X,\ Y$ with outcomes

$$[r] := \{1, ..., r\} \quad \text{and} \quad [c] := \{1, ..., c\}.$$

Recall that $p = (p_{ij})$, where

- $p_{ij} = P(X = i,\ Y = j)$,
- $p_{i+} = p_{i1} + \cdots + p_{ic}$,
- $p_{+j} = p_{1j} + \cdots + p_{rj}$.

Independent variables

Consider two random variables X, Y with outcomes

$$[r] := \{1, ..., r\} \quad \text{and} \quad [c] := \{1, ..., c\}.$$

Recall that $p = (p_{ij})$, where

- $p_{ij} = P(X = i,\ Y = j)$,
- $p_{i+} = p_{i1} + \cdots + p_{ic}$,
- $p_{+j} = p_{1j} + \cdots + p_{rj}$.

The random variables X and Y are **independent** if the joint probabilities factor as $p_{ij} = p_{i+}p_{+j}$ for all $i \in [r]$, $j \in [c]$. We denote it by $X \perp\!\!\!\perp Y$.

Proposition

$X \perp\!\!\!\perp Y \iff p = (p_{ij})$ *has rank 1.*

Proposition

$X \perp\!\!\!\perp Y \iff p = (p_{ij})$ *has rank 1.*

Proof.

($\Longrightarrow$) If $X \perp\!\!\!\perp Y$ then

$$(p_{ij}) = \begin{pmatrix} p_{11} & p_{12} & \dots & p_{1c} \\ \vdots & \vdots & \vdots & \vdots \\ p_{r1} & p_{r2} & \dots & p_{rc} \end{pmatrix} = \begin{pmatrix} p_{1+}p_{+1} & p_{1+}p_{+2} & \dots & p_{1+}p_{+c} \\ \vdots & \vdots & \vdots & \vdots \\ p_{r+}p_{+1} & p_{r+}p_{+2} & \dots & p_{r+}p_{+c} \end{pmatrix}$$

$$= \begin{pmatrix} p_{1+} \\ p_{2+} \\ \vdots \\ p_{r+} \end{pmatrix} \begin{pmatrix} p_{+1} & p_{+2} & \dots & p_{+c} \end{pmatrix}$$

Independence

Proposition

$X \perp\!\!\!\perp Y \iff p = (p_{ij})$ *has rank 1.*

Proof.

$(\Longrightarrow)$ If $X \perp\!\!\!\perp Y$ then

$$(p_{ij}) = \begin{pmatrix} p_{11} & p_{12} & \dots & p_{1c} \\ \vdots & \vdots & \vdots & \vdots \\ p_{r1} & p_{r2} & \dots & p_{rc} \end{pmatrix} = \begin{pmatrix} p_{1+}p_{+1} & p_{1+}p_{+2} & \dots & p_{1+}p_{+c} \\ \vdots & \vdots & \vdots & \vdots \\ p_{r+}p_{+1} & p_{r+}p_{+2} & \dots & p_{r+}p_{+c} \end{pmatrix}$$

$$= \begin{pmatrix} p_{1+} \\ p_{2+} \\ \vdots \\ p_{r+} \end{pmatrix} \begin{pmatrix} p_{+1} & p_{+2} & \dots & p_{+c} \end{pmatrix}$$

The map $(p_{i+}, p_{j+}) \mapsto p_{i+}p_{j+}$ is a Segre embedding.

Proposition

$X \perp\!\!\!\perp Y \iff p = (p_{ij})$ has rank 1.

Proof.

($\Longleftarrow$) If $\text{rank}(p) = 1$, then $p = \alpha b^T$, for $\alpha \in \mathbb{R}^r$, $b \in \mathbb{R}^c$. We choose α, b to be non-negative. Let $\alpha+ = \sum_{i=1}^r \alpha_i$, $b+ = \sum_{i=1}^c b_i$. Then $p_{ij} = \alpha_i b_j$, hence

$$p_{i+} = \alpha_i b_+, \;\; p_{+j} = \alpha_+ b_j, \;\; \text{and } \alpha_+ b_+ = b_+ \alpha_+ = 1.$$

Hence, $\quad p_{ij} = \alpha_i b_j = \alpha_i b_+ \alpha_+ b_j = p_{i+} p_{+j}, \quad \text{for all } i \in [r], j \in [c].$

■

Statistical model

Consider a $n-$sample of independent and identically distributed pairs of random variables

$$\begin{pmatrix} X^{(1)} \\ Y^{(1)} \end{pmatrix}, \begin{pmatrix} X^{(2)} \\ Y^{(2)} \end{pmatrix}, \ldots, \begin{pmatrix} X^{(n)} \\ Y^{(n)} \end{pmatrix},$$

$$P(X^{(k)} = i,\ Y^{(k)} = j) = p_{ij},\ \ \forall i\ \in\ [r],\ j\ \in\ [c],\ k\ \in\ [n].$$

The joint probability matrix $p = (p_{ij}) \in \Delta_{rc-1}$,

$$\Delta_{rc-1} = \{q \in \mathbb{R}^{r \times c}\ \mid\ q_{ij} \geq 0,\ \sum_{i=1}^{r} \sum_{j=1}^{c} q_{ij} = 1,\ \forall\ i, j\}.$$

A **statistical model** $\mathcal{M}$ is a subset of Δ_{rc-1}.

Independence model

The **independence model** for X and Y is the set

$$\mathcal{M}_{X \perp\!\!\!\perp Y} := \left\{ p \in \Delta_{rc-1} : \operatorname{rank}(p) = 1 \right\}.$$

$$\text{Then} \quad p = \begin{pmatrix} p_{11} & p_{12} & \dots & p_{1c} \\ p_{21} & p_{22} & \dots & p_{2c} \\ \vdots & \vdots & \vdots & \vdots \\ p_{r1} & p_{r2} & \dots & p_{rc} \end{pmatrix} \text{ and}$$

$$\operatorname{rank}(p) = 1 \iff p_{ij} p_{kl} - p_{il} p_{jk} = 0.$$

Independence model

The **independence model** for X and Y is the set

$$\mathcal{M}_{X \perp\!\!\!\perp Y} := \left\{ p \in \Delta_{rc-1} : \operatorname{rank}(p) = 1 \right\}.$$

$$\text{Then} \quad p = \begin{pmatrix} p_{11} & p_{12} & \dots & p_{1c} \\ p_{21} & p_{22} & \dots & p_{2c} \\ \vdots & \vdots & \vdots & \vdots \\ p_{r1} & p_{r2} & \dots & p_{rc} \end{pmatrix} \text{ and}$$

$$\operatorname{rank}(p) = 1 \iff p_{ij}p_{kl} - p_{il}p_{jk} = 0.$$

Let $S = \{x_{ij}x_{kl} - x_{il}x_{jk} \mid 1 \leq i < k \leq r,\ 1 \leq j < l \leq c\} \subset \mathbb{R}[x_{11}, x_{12}, \dots, x_{rc}]$.

$$V(S) = \{p \in \mathbb{R}^{r \times c} \mid \forall f(x_{11}, x_{12}, \dots, x_{rc}) \in S : f(p) = 0\}.$$

$$\mathcal{M}_{X \perp\!\!\!\perp Y} = \Delta_{rc-1} \cap V(S).$$

Independence model

The **independence model** for X and Y is the set

$$\mathcal{M}_{X \perp\!\!\!\perp Y} := \{p \in \Delta_{rc-1} : \text{rank}(p) = 1\}.$$

$$\text{Then} \quad p = \begin{pmatrix} p_{11} & p_{12} & \dots & p_{1c} \\ p_{21} & p_{22} & \dots & p_{2c} \\ \vdots & \vdots & \vdots & \vdots \\ p_{r1} & p_{r2} & \dots & p_{rc} \end{pmatrix} \text{ and}$$

$$\text{rank}(p) = 1 \iff p_{ij}p_{kl} - p_{il}p_{jk} = 0.$$

Let $S = \{x_{ij}x_{kl} - x_{il}x_{jk} \mid 1 \leq i < k \leq r,\ 1 \leq j < l \leq c\} \subset \mathbb{R}[x_{11}, x_{12}, \dots, x_{rc}]$.

$$V(S) = \{p \in \mathbb{R}^{r \times c} \mid \forall f(x_{11}, x_{12}, \dots, x_{rc}) \in S : f(p) = 0\}.$$

$$\mathcal{M}_{X \perp\!\!\!\perp Y} = \Delta_{rc-1} \cap V(S).$$ Segre variety

Contingency tables

Having $$\begin{pmatrix} X^{(1)} \\ Y^{(1)} \end{pmatrix}, \begin{pmatrix} X^{(2)} \\ Y^{(2)} \end{pmatrix}, \dots, \begin{pmatrix} X^{(n)} \\ Y^{(n)} \end{pmatrix},$$

we summarize the observations in a table of counts

$$U_{ij} = \sum_{k=1}^{n} 1_{\{X^{(k)}=i, Y^{(k)}=j\}}, \qquad i \in [r],\ j \in [c],$$

The table $U = (U_{ij})$ is a **two-way contingency table**.

The set of contingency tables that arise for sample size n is

$$\mathcal{T}(n) := \left\{ u \in \mathbb{N}^{r \times c} : \sum_{i=1}^{r} \sum_{j=1}^{c} u_{ij} = n \right\}.$$

	Death Penalty		
Race	Yes	No	Total
White	19	141	160
Black	17	149	166
Total	36	290	326

Classification of 326 homicide indictments in Florida in 1970s.

The contingency table shown above is represented as the table

$$u = \left(19, 141, 17, 149\right) \in \mathcal{T}(326).$$

$$\text{Also} \quad \left(14, 146, 22, 144\right), \left(20, 140, 16, 150\right), \ldots \in \mathcal{T}(326).$$

Hypothesis testing

Consider the **hypothesis testing** problem

$$H_0 : p \in \mathcal{M}_{X \perp\!\!\!\perp Y} \quad \text{(null hypothesis)}.$$

- ⋄ Chi-square test of independence.
- ⋄ Fisher's exact test.

Hypothesis testing

Consider the **hypothesis testing** problem

$$H_0 : p \in \mathcal{M}_{X \perp\!\!\!\perp Y} \quad \text{(null hypothesis).}$$

- Chi-square test of independence.
- Fisher's exact test.

Chi-square test of independence-Sketch

- If H_0 is true, then $p_{ij} = p_{i+}p_{+j}$, and the expected number of occurrences of the joint event $\{X = i, Y = j\}$ is $np_{i+}p_{+j}$.

Chi-square test of independence-Sketch

- If H_0 is true, then $p_{ij} = p_{i+}p_{+j}$, and the expected number of occurrences of the joint event $\{X = i, Y = j\}$ is $np_{i+}p_{+j}$.
- Use the empirical proportions

$$\hat{p}_{i+} = \frac{U_{i+}}{n} \quad \text{and} \quad \hat{p}_{+j} = \frac{U_{+j}}{n},$$

to estimate the marginal probabilities $p_{i+},\ p_{+j}$,

Chi-square test of independence-Sketch

- If H_0 is true, then $p_{ij} = p_{i+}p_{+j}$, and the expected number of occurrences of the joint event $\{X = i, Y = j\}$ is $np_{i+}p_{+j}$.
- Use the empirical proportions

$$\hat{p}_{i+} = \frac{U_{i+}}{n} \quad \text{and} \quad \hat{p}_{+j} = \frac{U_{+j}}{n},$$

  to estimate the marginal probabilities $p_{i+},\ p_{+j}$,
- estimate $np_{i+}p_{+j}$ by $\hat{u}_{ij} := n\hat{p}_{i+}\hat{p}_{+j}$,

Chi-square test of independence-Sketch

- If H_0 is true, then $p_{ij} = p_{i+}p_{+j}$, and the expected number of occurrences of the joint event $\{X = i, Y = j\}$ is $np_{i+}p_{+j}$.
- Use the empirical proportions

$$\hat{p}_{i+} = \frac{U_{i+}}{n} \quad \text{and} \quad \hat{p}_{+j} = \frac{U_{+j}}{n},$$

to estimate the marginal probabilities $p_{i+},\ p_{+j}$,

- estimate $np_{i+}p_{+j}$ by $\hat{u}_{ij} := n\hat{p}_{i+}\hat{p}_{+j}$,
- compute the **chi-square statistic**

$$X^2(U) := \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(U_{ij} - \hat{u}_{ij})^2}{\hat{u}_{ij}},$$

Chi-square test of independence-Sketch

- compute the probability

$$P(X^2(U) \geq X^2(u) \mid H_0 \text{ is true}).$$

Chi-square test of independence-Sketch

- compute the probability

$$P(X^2(U) \geq X^2(u) \mid H_0 \text{ is true}).$$

- This is known as the $p-$value of the statistical test. If the p-value is low (usually < 0.05) then we conclude that the null hypothesis was wrong. If the p-value is not low, then the chi-square test is inconclusive.

Example

	Death Penalty		
Race	Yes	No	Total
White	19	141	160
Black	17	149	166
Total	36	290	326

Classification of 326 homicide indictments in Florida in 1970s.

Starting with the matrix $u = (19, 17, 141, 149)$, $r = c = 2$, we can use R to compute that

$$p\text{-value} = 0.638.$$

The p-value is large; there is no evidence against the hypothesis.

Hypothesis testing

Consider the **hypothesis testing** problem

$$H_0 : p \in \mathcal{M}_{X \perp\!\!\!\perp Y} \quad \text{(null hypothesis).}$$

- Chi-square test of independence.
- Fisher's exact test.

Fisher's exact test

Proposition

Let $r = c = 2$. *If* $p = (p_{ij}) \in \mathcal{M}_{X \perp\!\!\!\perp Y}$ *and* $u \in \mathcal{T}(n)$, *then*

$$P(U_{11} = u_{11} \mid U_{1+} = u_{1+}, U_{+1} = u_{+1}) = \frac{\binom{u_{1+}}{u_{11}}\binom{n - u_{1+}}{u_{+1} - u_{11}}}{\binom{n}{u_{+1}}}$$

for $u_{11} \in \{max(0, u_{1+} + u_{+1} - n), ..., min(u_{1+}, u_{+1})\}$ *and zero otherwise.*

Fisher's exact test

Proposition

Let $r = c = 2$. If $p = (p_{ij}) \in \mathcal{M}_{X \perp\!\!\!\perp Y}$ and $u \in \mathcal{T}(n)$, then

$$P(U_{11} = u_{11} \mid U_{1+} = u_{1+}, U_{+1} = u_{+1}) = \frac{\begin{pmatrix} u_{1+} \\ u_{11} \end{pmatrix} \begin{pmatrix} n - u_{1+} \\ u_{+1} - u_{11} \end{pmatrix}}{\begin{pmatrix} n \\ u_{+1} \end{pmatrix}}$$

for $u_{11} \in \{max(0, u_{1+} + u_{+1} - n), ..., min(u_{1+}, u_{+1})\}$ and zero otherwise.

Proof. Recall that

$$P(U = u) = \begin{pmatrix} n \\ u \end{pmatrix} \prod_{i=1}^{r} \prod_{j=1}^{c} {p_{ij}}^{u_{ij}}$$
$$= \frac{n!}{u_{11}! u_{12}! ... u_{rc}!} \prod_{i=1}^{r} \prod_{j=1}^{c} {p_{ij}}^{u_{ij}} \ .$$

Fisher's exact test

Fix u_{1+} and u_{+1}. Then, as a function of u_{11}, the conditional probability

$$P(U_{11} = u_{11} \mid U_{1+} = u_{1+}, U_{+1} = u_{+1}) = \frac{P(U_{11} = u_{11}, U_{1+} = u_{1+}, U_{+1} = u_{+1})}{P(U_{1+} = u_{1+}, U_{+1} = u_{+1})}$$

$$= \frac{\binom{n}{u_{1+}}\binom{u_{1+}}{u_{11}}\binom{n-u_{1+}}{u_{+1}-u_{11}} p_{1+}^{u_{1+}} p_{2+}^{n-u_{1+}} p_{+1}^{u_{+1}} p_{+2}^{n-u_{+1}}}{\sum_{u_{11}} \binom{n}{u_{1+}}\binom{u_{1+}}{u_{11}}\binom{n-u_{1+}}{u_{+1}-u_{11}} p_{1+}^{u_{1+}} p_{2+}^{n-u_{1+}} p_{+1}^{u_{+1}} p_{+2}^{n-u_{+1}}}$$

$$= \frac{\binom{u_{1+}}{u_{11}}\binom{n-u_{1+}}{u_{+1}-u_{11}}}{\binom{n}{u_{+1}}}, \quad \text{since}$$

$$\sum_{u_{11}} \binom{u_{1+}}{u_{11}}\binom{n-u_{1+}}{u_{+1}-u_{11}} = \binom{n}{u_{1+}}.$$

Fisher's exact test

Let $u \in \mathcal{T}(n)$ be an observed $2 \times 2-$contingency table. Then

$$P(X^2(U) \geq X^2(u) \mid U_{1+} = u_{1+}, U_{+1} = u_{+1})$$

can be computed by summing over the probabilities

$$P(U_{11} = v_{11} \mid U_{1+} = v_{1+}, U_{+1} = v_{+1}) = \frac{\binom{u_{1+}}{v_{11}}\binom{n-u_{1+}}{u_{+1}-v_{11}}}{\binom{n}{u_{+1}}}$$

for all values $v_{11} \in \left\{max(0, u_{1+} + u_{+1} - n), ..., min(u_{1+}, u_{+1})\right\}$ such that $X^2(v) \geq X^2(u)$. In other words, the p-value is

$$\sum_{v_{11}} 1_{X^2(v) \geq X^2(u)} \frac{\binom{u_{1+}}{v_{11}}\binom{n-u_{1+}}{u_{+1}-v_{11}}}{\binom{n}{u_{+1}}}.$$

Example

Race	Death Penalty Yes	Death Penalty No	Total
White	19	141	160
Black	17	149	166
Total	36	290	326

Classification of 326 homicide indictments in Florida in 1970s.

For the contigency table $u = \Big(19, 141, 17, 149\Big) \in \mathcal{T}(326)$, the p-value is

$$\sum_{0 \leq v_{11} \leq 36} 1_{X^2(v) \geq X^2(u)} \frac{\binom{160}{v_{11}}\binom{166}{160 - v_{11}}}{\binom{326}{36}}.$$

General case

Let $X_1, ..., X_m$ be discrete random variables; X_i takes values in $[r_i]$.

General case

Let $X_1, ..., X_m$ be discrete random variables; X_i takes values in $[r_i]$.

Let $\mathcal{R} := \prod_{i=1}^{m}[r_i] = [r_1] \times [r_2] \times \cdots \times [r_m]$.

General case

Let $X_1, \ldots, X_m$ be discrete random variables; X_i takes values in $[r_i]$.

Let $\mathcal{R} := \prod_{i=1}^m [r_i] = [r_1] \times [r_2] \times \cdots \times [r_m]$.

Consider the joint probabilities

$$p_i = P(X_1 = i_1, \ldots, X_m = i_m), \quad i = (i_1, \ldots, i_m) \in \mathcal{R}.$$

The probability table $p = (p_i \mid i \in \mathcal{R})$ is in the $\#\mathcal{R} - 1-$ simplex

$$\Delta_{\mathcal{R}-1} = \{q \in \mathbb{R}^{\mathcal{R}} \mid q_i \geq 0, \sum_{i \in \mathcal{R}} q_i = 1\}.$$

General case

Let $X_1, ..., X_m$ be discrete random variables; X_i takes values in $[r_i]$.

Let $\mathcal{R} := \prod_{i=1}^{m}[r_i] = [r_1] \times [r_2] \times \cdots \times [r_m]$.

Consider the joint probabilities

$$p_i = P(X_1 = i_1, ..., X_m = i_m), \quad i = (i_1, ..., i_m) \in \mathcal{R}.$$

The probability table $p = (p_i \mid i \in \mathcal{R})$ is in the $\#\mathcal{R} - 1-$ simplex

$$\Delta_{\mathcal{R}-1} = \{q \in \mathbb{R}^{\mathcal{R}} \mid q_i \geq 0, \sum_{i \in \mathcal{R}} q_i = 1\}.$$

Let $\text{int}(\Delta_{\mathcal{R}-1})$ be the **interior** of $\Delta_{\mathcal{R}-1}$,

$$\text{int}(\Delta_{\mathcal{R}-1}) := \{p \in \Delta_{\mathcal{R}-1} \mid p > 0\}.$$

Log-linear model

Fix a matrix $A \in \mathbb{Z}^{d \times \mathcal{R}}$, whose columns all sum to the same value, $d \in \mathbb{N}$. The **log-linear model** (toric model) associated with A is the set of positive probability tables

$$\mathcal{M}_A = \left\{ p = (p_i) \in \text{int}(\Delta_{\mathcal{R}-1}) : \log p \in \text{rowspan}(A) \right\},$$

where $\text{rowspan}(A) = A^T b$ is the linear space spanned by the rows of A.

$$\mathcal{M}_A = \Delta_{\mathcal{R}-1} \cap V(I).$$

Observation

Let $p \in \mathcal{M}_{X \perp\!\!\!\perp Y}$. If p has all positive entries ($p \in \text{int}(\Delta_{\mathcal{R}-1})$) then

$$\begin{aligned} \log p &= (\log p_{1+}p_{+1}, \log p_{1+}p_{+2}, \log p_{2+}p_{+1}, \log p_{2+}p_{+2}) \\ &= (\log p_{1+} + \log p_{+1}, \log p_{1+} + \log p_{+2}, \log p_{2+} + \log p_{+1}, \log p_{2+} + \log p_{+2}) \\ &= \log p_{1+}(1,1,0,0) + \log p_{2+}(0,0,1,1) + \log p_{+1}(1,0,1,0) \\ &\quad + \log p_{+2}(0,1,0,1). \end{aligned}$$

Observation

Let $p \in \mathcal{M}_{X \perp\!\!\!\perp Y}$. If p has all positive entries ($p \in \text{int}(\Delta_{\mathcal{R}-1})$) then

$$\begin{aligned}
\log p &= (\log p_{1+}p_{+1}, \log p_{1+}p_{+2}, \log p_{2+}p_{+1}, \log p_{2+}p_{+2}) \\
&= (\log p_{1+} + \log p_{+1}, \log p_{1+} + \log p_{+2}, \log p_{2+} + \log p_{+1}, \log p_{2+} + \log p_{+2}) \\
&= \log p_{1+}(1,1,0,0) + \log p_{2+}(0,0,1,1) + \log p_{+1}(1,0,1,0) \\
&\quad + \log p_{+2}(0,1,0,1).
\end{aligned}$$

Thus $\log p \in \mathcal{M}_A$, where $A \in \mathbb{Z}^{4\times 4}$ is the matrix

$$A = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}.$$

Consider a vector of counts

$$U_i = \sum_{k=1}^{n} 1_{\left\{X_1^{(k)}=i_1,\ldots,X_m^{(k)}=i_m\right\}}, \quad i = (i_1, ..., i_m) \in \mathcal{R},$$

based on a random n-sample of independent and identically distributed vectors

$$\begin{pmatrix} X_1^{(1)} \\ \vdots \\ X_m^{(1)} \end{pmatrix}, \begin{pmatrix} X_1^{(2)} \\ \vdots \\ X_m^{(2)} \end{pmatrix}, \ldots, \begin{pmatrix} X_1^{(n)} \\ \vdots \\ X_m^{(n)} \end{pmatrix}.$$

The counts U_i now form an $m-$**way table** $U = (U_i) \in \mathbb{N}^{\mathcal{R}}$. Let

$$\mathcal{T}(n) = \left\{u \in \mathbb{N}^{\mathcal{R}} : \sum_{i \in \mathcal{R}} u_i = n\right\}.$$

Fiber

The vector Au is called the **minimal sufficient statistic(s)** for $\mathcal{M}_A$.
The set of tables

$$\mathcal{F}(u) = \left\{ v \in \mathbb{N}^{\mathcal{R}} : Av = Au \right\}$$

is called the **fiber** of the table $u \in \mathcal{T}(n)$ with respect to $\mathcal{M}_A$.

Fiber

The vector Au is called the **minimal sufficient statistic(s)** for $\mathcal{M}_A$.
The set of tables

$$\mathcal{F}(u) = \left\{ v \in \mathbb{N}^{\mathcal{R}} : Av = Au \right\}$$

is called the **fiber** of the table $u \in \mathcal{T}(n)$ with respect to $\mathcal{M}_A$.

We assumed that the columns of $A = (\alpha_{ji})$ sum to the same value, say $\sum_{j \in [d]} \alpha_{ji} = k$. Hence the tables in the fiber $\mathcal{F}(u)$ sum to n.

$$\begin{aligned} Av = Au \Rightarrow \sum_{i \in \mathcal{R}} \sum_{j \in [d]} \alpha_{ji} v_i &= \sum_{i \in \mathcal{R}} \sum_{j \in [d]} \alpha_{ji} u_i \\ \Rightarrow k \sum_{i \in \mathcal{R}} v_i &= k \sum_{i \in \mathcal{R}} u_i = kn. \end{aligned}$$

Proposition

If $p = e^{A^T b} \in \mathcal{M}_A$ and $u \in \mathcal{T}(n)$, then

$$P(U = u) = \binom{n}{u} e^{b^T (Au)}, \quad \text{where } \binom{n}{u} = \frac{n!}{\prod_{i \in \mathcal{R}} u_i!}.$$

Moreover, $P(U = u \mid AU = Au)$ does not depend on A or p.

Proof.

$$P(U = u) = \binom{n}{u} \prod_{i \in \mathcal{R}} p_i^{u_i} = \binom{n}{u} \prod_{i \in \mathcal{R}} e^{(A^T b)_i u_i} = \binom{n}{u} e^{b^T (Au)}$$

$$\text{and} \;\; P(AU = Au) = \sum_{v \in \mathcal{F}(u)} P(U = v)$$

$$= \sum_{v \in \mathcal{F}(u)} \binom{n}{v} e^{b^T (Av)} = e^{b^T (Au)} \sum_{v \in \mathcal{F}(u)} \binom{n}{v}.$$

Proposition

If $p = e^{A^T b} \in \mathcal{M}_A$ *and* $u \in \mathcal{T}(n)$, *then*

$$P(U = u) = \binom{n}{u} e^{b^T(Au)}, \quad \textit{where} \quad \binom{n}{u} = \frac{n!}{\prod_{i \in \mathcal{R}} u_i!}.$$

Moreover, $P(U = u \mid AU = Au)$ *does not depend on* A *or* p.

Proof.

Hence

$$P(U = u \mid AU = Au) = \frac{P(U = u)}{P(AU = Au)}$$

$$= \frac{e^{b^T(Au)} \binom{n}{u}}{e^{b^T(Au)} \sum_{v \in \mathcal{F}(u)} \binom{n}{v}} = \frac{\binom{n}{u}}{\sum_{v \in \mathcal{F}(u)} \binom{n}{v}}.$$

■

Generalised Fisher test

Consider the **hypothesis testing** problem

$$H_0 : p \in \mathcal{M}_A \quad \text{(null hypothesis)}.$$

Let

$$X^2(U) := \sum_{i \in \mathcal{R}} \frac{(U_i - \hat{u}_i)^2}{\hat{u}_i}.$$

We can generalize Fisher's exact test by computing the p-value

$$P(X^2(U) \geq X^2(u) \mid AU = Au)$$

$$= \sum_{w \in \mathcal{F}(u)} 1_{X^2(w) \geq X^2(u)} P(U = w \mid AU = Aw) = \frac{\sum_{w \in \mathcal{F}(u)} 1_{X^2(w) \geq X^2(u)} \cdot \binom{n}{w}}{\sum_{v \in \mathcal{F}(u)} \binom{n}{v}}.$$

Markov bases

- $A \in \mathbb{Z}^{d \times (\mathcal{R})}$ whose columns sum to the same row,

Markov bases

- $A \in \mathbb{Z}^{d \times (\mathcal{R})}$ whose columns sum to the same row,
- $\mathcal{M}_A = \{p \in \mathrm{int}(\Delta_{\mathcal{R}-1}) : \log p \in \mathrm{rowspan}(A)\}$,

Markov bases

- $A \in \mathbb{Z}^{d \times (\mathcal{R})}$ whose columns sum to the same row,
- $\mathcal{M}_A = \{p \in \text{int}(\Delta_{\mathcal{R}-1}) : \log p \in \text{rowspan}(A)\}$,
- $ker_{\mathbb{Z}}(A) = \left\{v \in \mathbb{N}^{\mathcal{R}} : Av = 0\right\}$ is the **integer kernel** of A,

Markov bases

- $A \in \mathbb{Z}^{d\times(\mathcal{R})}$ whose columns sum to the same row,
- $\mathcal{M}_A = \{p \in \mathrm{int}(\Delta_{\mathcal{R}-1}) : \log p \in \mathrm{rowspan}(A)\}$,
- $ker_{\mathbb{Z}}(A) = \left\{v \in \mathbb{N}^{\mathcal{R}} : Av = 0\right\}$ is the **integer kernel** of A,

Example

$$Av = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} v_{11} \\ v_{12} \\ v_{21} \\ v_{22} \end{pmatrix} = 0$$

$$\iff \begin{pmatrix} v_{11} + v_{12} \\ v_{21} + v_{22} \\ v_{11} + v_{21} \\ v_{12} + v_{22} \end{pmatrix} = \begin{pmatrix} v_{1+} \\ v_{2+} \\ v_{+1} \\ v_{+2} \end{pmatrix} = 0$$

$$\text{hence} \quad \ker_{\mathbb{Z}}(A) = \langle (1, -1, -1, 1) \rangle$$

Markov bases

- $A \in \mathbb{Z}^{d \times (\mathcal{R})}$ whose columns sum to the same row,
- $\mathcal{M}_A = \{p \in \text{int}(\Delta_{\mathcal{R}-1}) : \log p \in \text{rowspan}(A)\}$,
- $ker_{\mathbb{Z}}(A) = \left\{v \in \mathbb{N}^{\mathcal{R}} : Av = 0\right\}$ is the **integer kernel** of A,

Markov bases

- $A \in \mathbb{Z}^{d\times(\mathcal{R})}$ whose columns sum to the same row,
- $\mathcal{M}_A = \{p \in \text{int}(\Delta_{\mathcal{R}-1}) : \log p \in \text{rowspan}(A)\}$,
- $ker_{\mathbb{Z}}(A) = \left\{v \in \mathbb{N}^{\mathcal{R}} : Av = 0\right\}$ is the **integer kernel** of A,
- $\mathcal{T}(n) = \{u \in \mathbb{N}^{\mathcal{R}} : \sum_{i\in\mathcal{R}} u_i = n\}$,

Markov bases

- $A \in \mathbb{Z}^{d \times (\mathcal{R})}$ whose columns sum to the same row,
- $\mathcal{M}_A = \{p \in \text{int}(\Delta_{\mathcal{R}-1}) : \log p \in \text{rowspan}(A)\}$,
- $ker_{\mathbb{Z}}(A) = \left\{v \in \mathbb{N}^{\mathcal{R}} : Av = 0\right\}$ is the **integer kernel** of A,
- $\mathcal{T}(n) = \{u \in \mathbb{N}^{\mathcal{R}} : \sum_{i \in \mathcal{R}} u_i = n\}$,
- $\mathcal{F}(u) = \{v \in \mathbb{N}^{\mathcal{R}} : Av = Au\}$,

Markov bases

- $A \in \mathbb{Z}^{d \times (\mathcal{R})}$ whose columns sum to the same row,
- $\mathcal{M}_A = \{p \in \text{int}(\Delta_{\mathcal{R}-1}) : \log p \in \text{rowspan}(A)\}$,
- $ker_{\mathbb{Z}}(A) = \left\{v \in \mathbb{N}^{\mathcal{R}} : Av = 0\right\}$ is the **integer kernel** of A,
- $\mathcal{T}(n) = \{u \in \mathbb{N}^{\mathcal{R}} : \sum_{i \in \mathcal{R}} u_i = n\}$,
- $\mathcal{F}(u) = \{v \in \mathbb{N}^{\mathcal{R}} : Av = Au\}$,

Definition

A finite subset $\mathcal{B} \subset ker_{\mathbb{Z}}(A)$ is a **Markov basis** for $\mathcal{M}_A$ if for all $u \in \mathcal{T}(n)$ and all pairs $v, v' \in \mathcal{F}(u)$, there exists a sequence $u_1, .., u_L \in \mathcal{B}$, such that

$$v' = v + \sum_{k=1}^{L} u_k \quad \& \quad v + \sum_{k=1}^{m} u_k \geq 0 \quad \text{for all } \; m = 1, ..., L.$$

The elements of the Markov basis are called moves.

A finite subset $\mathcal{B} \subset ker_{\mathbb{Z}}(A)$ is a **Markov basis** for $\mathcal{M}_A$ if for all $u \in \mathcal{T}(n)$ and all pairs $v, v' \in \mathcal{F}(u)$, there exists a sequence $u_1, .., u_L \in \mathcal{B}$, such that

$$v' = v + \sum_{k=1}^{L} u_k \quad \& \quad v + \sum_{k=1}^{m} u_k \geq 0 \quad \text{for all} \;\; m = 1, ..., L.$$

The elements of the Markov basis are called moves.

Example

$$\text{If } v = \Big(14, 146, 22, 144\Big), \; v' = \Big(20, 140, 16, 150\Big) \in \mathcal{F}(u) \text{ and}$$

$$\Big(1, -1, -1, 1\Big) \in \mathcal{B}, \; \text{then}$$

$$v' = v + 4\Big(1, -1, -1, 1\Big) \;\; \& \;\; v + m\Big(1, -1, -1, 1\Big) \geq 0, \quad m \in [4].$$

Example

$$A = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \Rightarrow \ker_{\mathbb{Z}}(A) = \langle (1, -1, -1, 1) \rangle$$

$$u = \left(u_{11}, u_{12}, u_{21}, u_{22}\right) = \left(19, 141, 17, 149\right) \in \mathcal{T}(326) \Rightarrow$$

$$\mathcal{F}(u) = \left\{ \left(19 + m, 141 - m, 17 - m, 149 + m\right) \mid -19 \leq m \leq 17 \in \mathbb{N}^4 \right\}$$

Example

$$\ker_{\mathbb{Z}}(A) = \langle(1,-1,-1,1)\rangle,$$

$$\mathcal{F}(u) = \Big\{ \Big(19+m, 141-m, 17-m, 149+m\Big) \mid -19 \le m \le 17 \in \mathbb{N}^4 \Big\}.$$

If $v,\ v' \in \mathcal{F}(u),\ \mathcal{B} = \Big\{ \pm \Big(1,-1,-1,1\Big) \Big\}$, then either

$$v' = v + L\Big(1,-1,-1,1\Big) \ \ \&\ \ v + m\Big(1,-1,-1,1\Big) \ge 0, \quad m = 1,\ldots,L, \text{ or}$$

$$v' = v + L\Big(-1,1,1,-1\Big) \ \ \&\ \ v + m\Big(-1,1,1,-1\Big) \ge 0, \quad m = 1,\ldots,L,$$

hence $\mathcal{B}$ is a Markov basis for $\mathcal{M}_A$.

Example

$$\ker_{\mathbb{Z}}(A) = \langle (1, -1, -1, 1) \rangle,$$

$$\mathcal{F}(u) = \left\{ \left(19 + m, 141 - m, 17 - m, 149 + m\right) \mid -19 \leq m \leq 17 \in \mathbb{N}^4 \right\}.$$

If $v,\ v' \in \mathcal{F}(u),\ \mathcal{B} = \left\{ \pm \left(1, -1, -1, 1\right) \right\},$ then either

$$v' = v + L\left(1, -1, -1, 1\right) \ \ \& \ \ v + m\left(1, -1, -1, 1\right) \geq 0, \ \ m = 1, ..., L, \text{ or}$$

$$v' = v + L\left(-1, 1, 1, -1\right) \ \ \& \ \ v + m\left(-1, 1, 1, -1\right) \geq 0, \ \ m = 1, ..., L,$$

hence $\mathcal{B}$ is a Markov basis for $\mathcal{M}_A$.

- Is it unique?

Example

$$\ker_{\mathbb{Z}}(A) = \langle (1,-1,-1,1) \rangle,$$

$$\mathcal{F}(u) = \Big\{ \Big(19+m, 141-m, 17-m, 149+m\Big) \mid -19 \le m \le 17 \in \mathbb{N}^4 \Big\}.$$

If $v,\ v' \in \mathcal{F}(u),\ \mathcal{B} = \Big\{ \pm \Big(1,-1,-1,1\Big) \Big\}$, then either

$$v' = v + L\Big(1,-1,-1,1\Big) \ \ \&\ \ v + m\Big(1,-1,-1,1\Big)\, 0, \quad m = 1, \dots, L, \text{ or}$$

$$v' = v + L\Big(-1,1,1,-1\Big) \ \ \&\ \ v + m\Big(-1,1,1,-1\Big)\, 0, \quad m = 1, \dots, L,$$

hence $\mathcal{B}$ is a Markov basis for $\mathcal{M}_A$.

- Is it unique?
- Is it "minimal"?

	TEMPERATURE			
MEDICINE	UP	DOWN	STABLE	TOTAL
CORTICOID	160	180	23	363
ANTIBIOTIC	82	91	15	188
PLACEBO	54	60	15	129
TOTAL	296	331	53	680

Let $A \in \mathbb{Z}^{6\times 9}$, be the matrix whose columns sum to 2:

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \end{pmatrix}.$$

	TEMPERATURE			
MEDICINE	UP	DOWN	STABLE	TOTAL
CORTICOID	160	180	23	363
ANTIBIOTIC	82	91	15	188
PLACEBO	54	60	15	129
TOTAL	296	331	53	680

Let $A \in \mathbb{Z}^{6\times 9}$, be the matrix whose columns sum to 2:

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \end{pmatrix}.$$

$$\mathcal{B} = \pm\left\{(1,-1,0,0,0,0,0,1,-1),(0,0,1,-1,1,0,0,-1,0),(1,0,0,1,0,-1,-1,0,0)\right\}$$

Random walk in $\mathcal{F}(u)$

Starting with a contingency table $u \in \mathcal{T}(n)$ and a Markov basis $\mathcal{B}$ for the model $\mathcal{M}_A$ we can use Metropolis-Hastings algorithm to compute a sequence $(X^2(v_t))_{t=1}^{\infty}$, where $v_t \in \mathcal{F}(u)$.

Theorem

With probability one, the output sequence $(X^2(v_t))_{t=1}^{\infty}$ satisfies

$$\lim_{M\to\infty} \frac{1}{M} \sum_{t=1}^{M} 1_{\{X^2(v_t) \geq X^2(u)\}} = P(X^2(U) \geq X^2(u) \mid AU = Au),$$

which is the p-value in the generalized Fisher's exact test!

Metropolis-Hastings algorithm

Input: $u \in \mathcal{T}(n)$, $\mathcal{B}$ Markov basis for the model $\mathcal{M}_A$.
Output: A sequence $(X^2(v_t))_{t=1}^{\infty}$ for tables v_t in the fiber $\mathcal{F}(u)$.
Step 1: Initialize $v_1 = u$.
Step 2: For $t = 1, 2, \ldots$ repeat the following steps:
(i) Select uniformly at random a move $u_t \in \mathcal{B}$.
(ii) If $\min(v_t + u_t) < 0$, then set $v_{t+1} = v_t$, else set

$$v_{t+1} = \begin{cases} v_t + u_t \\ v_t \end{cases} \quad \text{with probability} \quad \begin{cases} q \\ 1 - q \end{cases},$$

where

$$q = \min\left\{1, \frac{P(U = v_t + u_t \mid AU = Au)}{P(U = v_t \mid AU = Au)}\right\}.$$

(iii) Compute $X^2(v_t)$.

Metropolis-Hastings algorithm

Input: $u \in \mathcal{T}(n)$, $\mathcal{B}$ Markov basis for the model $\mathcal{M}_A$.
Output: A sequence $(X^2(v_t))_{t=1}^{\infty}$ for tables v_t in the fiber $\mathcal{F}(u)$.
Step 1: Initialize $v_1 = u$.
Step 2: For $t = 1, 2, \ldots$ repeat the following steps:

(i) Select uniformly at random a move $u_t \in \mathcal{B}$.

(ii) If $\min(v_t + u_t) < 0$, then set $v_{t+1} = v_t$, else set

$$v_{t+1} = \begin{cases} v_t + u_t \\ v_t \end{cases} \quad \text{with probability} \quad \begin{cases} q \\ 1 - q \end{cases},$$

where

$$q = \min\left\{1, \frac{P(U = v_t + u_t \mid AU = Au)}{P(U = v_t \mid AU = Au)}\right\}.$$

(iii) Compute $X^2(v_t)$.

$$P(U = u \mid AU = Au) = \frac{\binom{n}{u}}{\sum_{v \in \mathcal{F}(u)} \binom{n}{v}}.$$