Fisher’s Exact Test

Kaie Kubjas, 28.10.2020

Agenda

- Last time: Maximum likelihood estimation
- This time: Hypothesis testing
- Does the unknown distribution, for which we have i.i.d. data, belong to a given model?
- <u>Discrete exponential families</u>
- Diaconis and Sturmfels (1998): “Algebraic algorithms for sampling from conditional distributions” - the beginning of algebraic statistics

Murder accusations in Florida

The following contingency table presents a classification of 326 murder accusations in Florida in the 1970s:

race\death penalty	yes	no	total
white	19	141	160
black	17	149	166
total	36	290	326

We would like to know whether the charge of death penalty was independent of the race. [Poll]

NB! We switch between <u>contingency tables</u> and <u>vectors of counts</u> as convenient.

Murder accusations in Florida

- Two discrete random variables:
  - X for defendant's race
  - Y for death penalty
- They both have two possible outcomes:
  - {white, black}
  - {yes,no}

Discrete exponential families

Fix $A = (a_{jx})_{j\in[k],x\in[r]} \in \mathbb{Z}^{k\times r}$ and $h \in \mathbb{R}^r_{>0}$.

Def: The discrete exponential family $\mathscr{M}_{A,h}$ consists of distributions

$$p_\theta(x) = \frac{1}{Z(\theta)} h_x \prod_j \theta_j^{a_{jx}} \text{ where } Z(\theta) = \sum_{x\in\mathscr{X}} h_x \prod_j \theta_j^{a_{jx}}.$$

The monomials $\prod_j \theta_j^{a_{jx}}$ correspond to columns of the matrix A.

Example: Let $A = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 3 & 2 & 1 & 0 \end{pmatrix}$ and $h = \mathbf{1}$. Then

$$p_\theta = \frac{1}{Z(\theta)} \left(\theta_2^3, \theta_1\theta_2^2, \theta_1^2\theta_2, \theta_1^3\right) \text{ where } Z(\theta) = \theta_2^3 + \theta_1\theta_2^2 + \theta_1^2\theta_2 + \theta_1^3.$$

Murder accusations in Florida

Poll: What is the matrix A and the vector h for the independence model of two binary random variables?

- Recall that a the parametrization of the independence model is given by

$$p_{ij} = \alpha_i \beta_j,$$

where $i \in [2], j \in [2]$ and α_i, β_j are independent parameters.

Answer: $A = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$ and $h = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$

Hypothesis testing

- A discrete exponential family $\mathscr{M}_{A,h} \subseteq \Delta_{r-1}$
- I.i.d. data $X^{(1)}, \ldots, X^{(n)} \in [r]$ from a distribution $p \in \text{int}(\Delta_{r-1})$
- We would like to test the hypothesis

$$H_0 : p \in \mathscr{M}_{A,h} \text{ versus } H_1 : p \notin \mathscr{M}_{A,h}$$

Hypothesis testing

- We would like to test the hypothesis

$$H_0 : p \in \mathscr{M}_{A,h} \text{ versus } H_1 : p \notin \mathscr{M}_{A,h}$$

- Hypothesis tests often use p-values
- p-value is the probability of obtaining a dataset that is at least as extreme as the observed dataset assuming that the null hypothesis H_0 is correct
- If <u>p-value is small</u> (e.g. less than 0.05), then the <u>null hypothesis is rejected</u>

Pearson's X^2 statistic

Def: Let X be a random vector taking values in a set $\mathscr{X}$. A statistic is a function from $\mathscr{X}$ to $\mathbb{R}^k$ for some $k \in \mathbb{N}$.

- Let $T : \mathbb{N}^r \to \mathbb{R}$ be a statistic which is zero if and only if $u/n \in \mathscr{M}_{A,h}$ and increases away from $\mathscr{M}_{A,h}$
- p-value: $Pr[T(v) > T(u) \,|\, H_0]$ where $v \in \mathbb{N}^r, ||v|| = n$.

Pearson's χ^2 statistic

- Pearson's χ^2 statistic: $X_n^2(u) = \sum_{j=1}^{r} \frac{(u_j - \hat{u}_j)^2}{\hat{u}_j}$, where $\hat{u} = n\hat{p}$ is the MLE of counts
- Pearson's χ^2 statistic converges to chi-square distribution with $df = r - 1 - \dim \mathscr{M}_{A,h}$ degrees of freedom
- If the sample size is small, it is not reasonable to consider sample size tending to infinity

Question 1

- Question 1: For which datasets are results of this paper useful?
- Answer:
  - I.i.d. samples from a discrete distribution
  - Small sample sizes: some of the entries of the contingency table are smaller than 5

Main idea

- $A = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$ and $u = \begin{pmatrix} 19 \\ 141 \\ 17 \\ 290 \end{pmatrix}$
- We consider all $v \in \mathbb{N}^4$ such that $Av = Au$
- Likelihood functions give a distribution on all such v
- What is the probability of observing a dataset as extreme as u?

Fibers

Def: Let $A \in \mathbb{Z}^{k \times r}$ and let $u \in \mathbb{N}^r$. The set of tables

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

is called the fiber of a contingency table u with respect to A.

Poll: Let $A = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$ and $u = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$. Which of the following vectors belong to the fiber $\mathcal{F}(u)$?

Question 2

- Question 2: What is the set $\mathscr{F}(u)$ in the case of the independence model?
- Answer: It consists of all the contingency tables that have the same row and column sums as u.

Likelihood function

- The likelihood function: $L(v \mid \theta) = \binom{n}{v} h^v \theta^{Av} Z(\theta)^{-n}$,

where $\binom{n}{u} = \dfrac{n!}{u_1! \cdots u_r!}$ is the multinomial coefficient.

- The likelihood function: $L(v \mid v \in \mathcal{F}(u), \theta) = \dfrac{\binom{n}{v} h^v \theta^{Av} Z(\theta)^{-n}}{\sum_{v \in \mathcal{F}(u)} \binom{n}{v} h^v \theta^{Av} Z(\theta)^{-n}}$

Statistic

Def: For a parametric statistical model $\mathscr{M}_{\Theta}$, a statistic T is sufficient if the probability density function or probability mass function factorizes as $f_\theta(x) = h(x)g(T(x), \theta)$.

Equivalently, a statistic T is sufficient if

$$P(X = x \mid T(X) = t, \theta) = P(X = x \mid T(X) = t).$$

A statistic T is minimal sufficient if every other sufficient statistics is a function of T.

Maximum likelihood

$$L(v \,|\, v \in \mathscr{F}(u), \theta) = \frac{\binom{n}{v} h^v \theta^{Av} Z(\theta)^{-n}}{\sum_{v \in \mathscr{F}(u)} \binom{n}{v} h^v \theta^{Av} Z(\theta)^{-n}}$$

The vector Au is the minimal sufficient statistic for the model $\mathscr{M}_{A,h}$.

All the terms involving θ cancel out, because $Av = Au$ for all $v \in \mathscr{F}(u)$:

$$L(v \,|\, v \in \mathscr{F}(u), \theta) = L(v \,|\, v \in \mathscr{F}(u)) = \frac{\binom{n}{v} h^v}{\sum_{v \in \mathscr{F}(u)} \binom{n}{v} h^v}$$

Distribution on the fiber

- The resulting distribution on the fiber $\mathscr{F}(u)$ where $P(v) \propto \binom{n}{v} h^v$ is called the generalized hypergeometric distribution
- To compute the p-value, we have to compute or approximate the sum

$$\frac{1}{\#\mathscr{F}(u)} \sum_{v \in \mathscr{F}(u)} 1_{T(v) \geq T(u)} L(v \,|\, v \in \mathscr{F}(u))$$

Murder accusations in Florida

- The fiber of the death penalty versus race table consists of 37 contingency tables

$$\begin{pmatrix} 0 & 160 \\ 36 & 130 \end{pmatrix}, \ldots, \begin{pmatrix} 18 & 142 \\ 18 & 148 \end{pmatrix}, \begin{pmatrix} 19 & 141 \\ 17 & 149 \end{pmatrix}, \begin{pmatrix} 20 & 140 \\ 16 & 150 \end{pmatrix}, \ldots, \begin{pmatrix} 36 & 124 \\ 0 & 166 \end{pmatrix}.$$

- The distribution on the fibers: $P(v) \propto \binom{n}{v} h^v$

- Since $h = \mathbf{1}$, then $P(v) \propto \binom{n}{v}$

-
```
In[1]:= Factorial[326] / (Factorial[19] * Factorial[141] * Factorial[17] * Factorial[149])
Out[1]= 775 268 042 602 097 147 736 537 522 819 203 932 553 604 398 847 142 652 948 456 333 859 390 634 184 127
        834 390 848 468 427 314 216 890 667 073 886 390 493 353 978 384 896 165 621 076 800 000
```

Markov bases

Markov bases

- In general, even enumerating the fiber is too difficult
- Alternative: Generate random samples from the fiber $\mathscr{F}(u)$ to get an estimate of the p-value

Markov basis

Def: Let $A \in \mathbb{Z}^{k\times r}$. Let $\ker_{\mathbb{Z}}(A) = \{v \in \mathbb{Z}^r : Av = 0\}$ be the integer kernel of A. A finite subset $\mathscr{B} \subset \ker_{\mathbb{Z}}(A)$ is a Markov basis for A if for all $u \in \mathbb{N}^n$ and all $u' \in \mathscr{F}(u)$ there exists a sequence $v_1, \ldots, v_L \in \mathscr{B}$ such that

$$u' = u + \sum_{k=1}^{L} v_k \text{ and } u + \sum_{k=1}^{L} v_k \geq 0 \text{ for all } l = 1,\ldots,L\,.$$

The elements of the Markov basis are called moves.

Graph interpretation

Markov basis example

Poll: What is a Markov basis for $A = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$?

Metropolis-Hastings

- Input: A contingency table $u \in \mathbb{N}^r$ and a Markov basis $\mathscr{B}$ for A.
- Output: A sequence of tables $u_1, u_2, \ldots \in \mathscr{F}(u)$.
- Step 1: Initialize $u_1 = u$.
- Step 2: For $t = 1,2,\ldots$ repeat the following steps:
    - Select uniformly at random a move $v_t \in \pm\mathscr{B}$.
    - If $\min(u_t + v_t) < 0$, then set $u_{t+1} = u_t$, else set

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

where $q = \min\left\{1, \dfrac{p(u_t + v_t)}{p(u_t)}\right\}$.

- Output the sequence $u_1, u_2, \ldots$.

Metropolis-Hastings

- The sequence of tables produced by Metropolis-Hastings eventually converges to a random sample from the desired distribution p.
- These samples can be used to compute the p-value.
- A major unsolved research problem: When is a sample closed to the desired distribution?
- algstat package in R

How to find a Markov basis?

Monomial parametrization map

Def: Let $A \in \mathbb{Z}^{k \times r}$ and $h \in \mathbb{R}^r_{>0}$. The monomial map associated to this data is the rational map

$$\phi^{A,h} : \mathbb{R}^k \to \mathbb{R}^r, \text{ where } \phi_j^{A,h} = h_j \prod_{i=1}^{k} \theta_i^{a_{ij}}.$$

NB! The normalizing constant $Z(\theta)$ is removed.

Example: Let $A = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 3 & 2 & 1 & 0 \end{pmatrix}$. The monomial map is $\phi^A : \mathbb{R}^2 \to \mathbb{R}^4$ is given by

$$(\theta_1, \theta_2) \mapsto (\theta_2^3, \theta_1\theta_2^2, \theta_1^2\theta_2, \theta_1^3).$$

Toric ideal

<u>Def:</u> Let $A \in \mathbb{Z}^{k \times r}$ and $h \in \mathbb{R}^r_{>0}$. The ideal

$$I_{A,h} := I(\phi^{A,h}(\mathbb{R}^k)) \subseteq \mathbb{R}[p]$$

is called the toric ideal associated to the pair A and h.

- If $h = \mathbf{1}$, then we denote $I_A := I_{A,1}$.
- Generators for the ideal $I_{A,h}$ are obtained from generators of the ideal I_A.

Toric ideal

Prop: Let $A \in \mathbb{Z}^{k\times r}$ and $h \in \mathbb{R}^r_{>0}$. Then

$$I_A = \langle p^u - p^{u'} : u, v \in \mathbb{N}^r \text{ and } Au = Au' \rangle.$$

Example: Let $A = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 3 & 2 & 1 & 0 \end{pmatrix}$. The monomial map is $\phi^A : \mathbb{R}^2 \to \mathbb{R}^4$ is given by

$$(\theta_1, \theta_2) \mapsto (\theta_2^3, \theta_1\theta_2^2, \theta_1^2, \theta_2, \theta_1^3).$$

The toric ideal is

$$I_A = \langle p_1p_3 - p_2^2, p_1p_4 - p_2p_3, p_2p_4 - p_3^2 \rangle.$$

Toric ideal

- Toric ideal is related to $\ker_{\mathbb{Z}}(A) = \{v \in \mathbb{Z}^r : Av = 0\}$
- $v \in \ker_{\mathbb{Z}}(A)$ can be written as $v = v^+ - v^-$
  - $v_j^+ = \max(v_j, 0)$
  - $v_j^- = -\max(-v_j, 0)$
- To $v \in \ker_{\mathbb{Z}}(A)$ associate $p^{v^+} - p^{v^-} \in I_A$
- Conversely, if $p^u - p^{u'} \in I_A$, then $u - u' \in \ker_{\mathbb{Z}}(A)$.

Fundamental theorem of Markov bases

Theorem: A subset $\mathscr{B}$ of $\ker_{\mathbb{Z}}(A)$ is a Markov basis if and only if the corresponding set of binomials $\{p^{b^+} - p^{b^-} : b \in \mathscr{B}\}$ generates the toric ideal I_A.

- Markov bases exist.
- An algebraic method for computing Markov bases.

```
Macaulay2, version 1.16.0.2
-- storing configuration for package FourTiTwo in /home/m2user/.Macaulay2/init-FourTiTwo.m2
-- storing configuration for package Topcom in /home/m2user/.Macaulay2/init-Topcom.
with packages: ConwayPolynomials, Elimination, IntegralClosure, InverseSystems, LLLBases, MinimalPrir

i1 : needsPackage "FourTiTwo"

o1 = FourTiTwo

o1 : Package

i2 : A = matrix{{1,1,0,0},{0,0,1,1},{1,0,1,0},{0,1,0,1}}
```

$$o2 = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

o2 : Matrix $\mathbb{Z}^4 \longleftarrow \mathbb{Z}^4$

```
i3 : B=toricMarkov A
```

$$o3 = \begin{pmatrix} 1 & -1 & -1 & 1 \end{pmatrix}$$

o3 : Matrix $\mathbb{Z}^1 \longleftarrow \mathbb{Z}^4$

```
i4 : R=QQ[p1,p2,p3,p4]
```

o4 = R

o4 : PolynomialRing

```
i5 : I=toBinomial(B,R)
```

o5 = ideal$(-p2\,p3 + p1\,p4)$

o5 : Ideal of R

copy to editor

Question 3

- Question 3: Do you recognize the ideal in Theorem 3.1 from lectures?
- Answer: It is the toric ideal $I_{A,h} := I(\phi^{A,h}(\mathbb{R}^k)) \subseteq \mathbb{R}[p]$.

Conclusion

- Hypothesis testing for discrete exponential families
- We want to compute the p-value
- Focus on small sample size
- We can compute the p-value on the fiber
- Markov bases together with Metropolis-Hastings allow to sample from a fiber
- Algebraic geometry is used for computing Markov bases

Group work

- No lectures during the next three weeks
- First two weeks: Make a presentation with the first group
- Third week: Present in a new group
- Exercise sessions / office hours take place as usual
- The last two lectures will be on graphical models